CHARACTERIZATION OF THE TYPE I-E CRISPR SYSTEM OF *T. FUSCA* AND ITS APPLICATION TO GENOME EDITING

A Dissertation

Presented to the Faculty of the Graduate School

of Cornell University

in Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

By

Adam Emin Dolan

December 2018

Characterization of the Type IE CRISPR System of *T. fusca* and Its Application to Genome Editing

Adam Emin Dolan

Cornell University, 2018

ABSTRACT

Genome editing techniques and platforms for nucleic acid targeting have been revolutionized by the discovery and technological adaptation of bacterial immune systems called CRISPR systems. These immune systems can specifically recognize, bind to, and cleave substrate nucleic acids to prevent phage infection for the bacterial strains which contain them. In genetic manipulation technologies, CRISPR systems allow for a high degree of targeting versatility along with simplified experimental design compared to other platforms. While Cas9 technologies continue to prove successful, other CRISPR systems exist which deserve investigation for their unique biochemical properties.

In this thesis, I first provide background on the mechanisms of CRISPR biochemistry. I focus primarily on the Type I-E CRISPR systems of *E. coli* and *T. fusca* (Cascade and Cas3) and make analogies to the well-characterized Type II-A system of *S. pyogenes* (Cas9). I provide some background on the discovery of CRISPR systems and touch on the organization and diversity of these immune systems in nature. I then also discuss some current trends in CRISPR technology, and since most of these technologies utilize the Cas9 platform, I discuss Cas9 in more detail.

Then I explore the mechanism of interaction between Cas3 and Cascade by mutational interface perturbation. The linker-helix region of Cas3 has been established to be important for the interaction

between Cascade and Cas3. However, the precise mechanism was difficult to determine without a detailed structure of Cascade interacting with Cas3. While there now exists a Cascade-Cas3 high-resolution Cryo-EM structure, my mutational analysis confirms that the interface at the linker helix is extensive and redundant. However, my analysis also suggests a more nuanced and not completely understood mechanism, since there were several mutations which caused observed binding defects which did not participate in the Cascade-Cas3 interface in the Cryo-EM structure.

Next I describe a genome editing technology based on the Type I-E CRISPR system of *T. fusca* which we applied in a human embryonic stem cell (hESC) line. When applied to hESCs, the Type I-E system introduces a spectrum of deletions that are caused by Cas3 and Cascade. These deletions have non-determinate start and stop sites, ranging from several hundred bases to many kilobases deleted. These deletions did not start at the targeting site, implying that some distance of Cas3 translocation is required before double-strand breaks are induced.

Finally, I describe efforts to elucidate the mechanism of Cas3-induced double-strand breaks. The observation that DSBs form distal to the target site in Cascade/Cas3-mediated genome editing implies a new mechanism of Cas3 activity. I present speculation and preliminary data which suggests that Cas3 dimerization is a plausible hypothesis to explain this behavior and I provide a road-map for investigation.

BIOGRAPHICAL SKETCH

Adam Dolan grew up in Nutley, NJ, along with his brother John. Throughout his elementary, middle, and high school education, Adam was fortunate enough to have teachers and a curriculum which fostered interest in many topics, including but not limited to science. His parents also cultivated Adam's enjoyment of travel which matured into an appreciation of diversity, and when Adam graduated from high school, he joined the inaugural class of New York University Abu Dhabi along with 150 other students from around the world. At NYUAD, Adam double majored in Biology and Physics while having the opportunity to explore philosophy and art history. While at NYUAD, Adam met and fell in love with Florencia Schlamp. Together, they applied to graduate schools and luckily both were accepted to PhD programs at Cornell University. Adam joined the lab of Prof. Ailong Ke. There, Adam wanted to work on CRISPR technology development, so he initiated a project to apply the Type I-E CRISPR system for genome editing. Adam lives in Ithaca NY, along with Florencia and their cat Numi.

ACKNOWLEDGEMENTS

I would like to acknowledge the people and institutions which have helped me to complete this thesis.

First and foremost, I thank Florencia for her empathy, for her love, and most of all for her support.

I thank my parents and my brother for the support they have shown me through my whole life.

I thank New York University Abu Dhabi for the education I received there and for the friendships that attending that institution allowed me to form and maintain. Juan Beltran, Attilio Rigotti, Zach Ross, Eric Johnson, Amelia Kahn, Katy Blumer, and others have been a great comfort to me and a source of support that I value dearly. They are as brothers and sisters to me and their examples are the compass by which I try to guide my own life.

I thank my committee, especially Professor Eric Alani, for their assistance and sage advice.

I thank Professor Ailong Ke and everyone in the lab for their support. I thank Drs. Ian Price and Robert Hayes for being the heart of the lab and for their friendship. I thank Robert Battaglia for his conversation and for being a peer that I respect and admire. I thank Dr. Yibei Xiao for his help regarding experiments and for his willingness to teach me at the start of my time in the lab. I thank Dr. Jagat Budhathoki for his guidance in learning single molecule techniques.

I thank Cornell University and the city of Ithaca for providing an environment in which I am glad to have participated. The friendships I have made here have ensured that Ithaca will hold a place of fondness in my heart for the rest of my life.

And last but not least, I thank my cat, Numi. Without her love and cuddles, this difficult endeavor would have been made much harder.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1 – INTRODUCTION

1.1 Introduction and general principles of CRISPR

The struggle between host genomes and invading genomes has been a defining drama of evolutionary history. At its core, the drive for a genome to defend itself from invaders is central to life itself on the most basic level. The need to evade, neutralize, or incapacitate invaders has caused a vast array of defenses to spring up in all branches of life from simple bacterial restriction enzymes to the complex immune responses in mammals. As the effectiveness of these systems improve, viruses and other invaders acquire their own mechanisms of escape and resistance to systems of immunity. Thus, there is a strong advantage to any immune system which can adapt to neutralize a diversity of invaders while also limiting the possibility of escape or resistance in viruses.

Adaptive immune systems are those which can sense an invader, affect immunity to the invader, and afterwards retain a memory of the infection – making a future infection less costly or impossible. These principles are applied in the adaptive immune response of mammals where genetic memory of antibody structural recognition of an invader is retained in Memory B cells and can be called up to neutralize a future infection (1). Likewise, bacteria have an adaptive immune response with strong analogies to the mammalian case (2). In this case, instead of protein-protein interactions mediating the recognition of viral or bacterial surface proteins, in bacteria, RNA-guided recognition directly recognizes the genomes of invading phages (3-6).

These bacterial adaptive immune systems are known as CRISPR (Clustered Regularly Interspaced Short Palindromic Repeat) systems, named so after their hallmark genetic element, the CRISPR array (7). These arrays are a genetic memory of previous phage infections and are passed down from mother cell to daughter cell within a population (8). These arrays are themselves comprised of the "Repeat" element – a sequence that, when transcribed into RNA often forms a structured region – and "Spacers"

(Figure 1.1a). Spacers are sequences that are derived from invader genomes and represent the library of nucleic acid sequences targeted by a given CRISPR array (9). CRISPR arrays have no defined maximum size, though there has been analysis that suggest there is an optimal CRISPR array size which is dependent on the environmental diversity of phage (10, 11). In contrast, the minimally defined CRISPR array that can affect immunity is a single spacer flanked by two repeats (12).

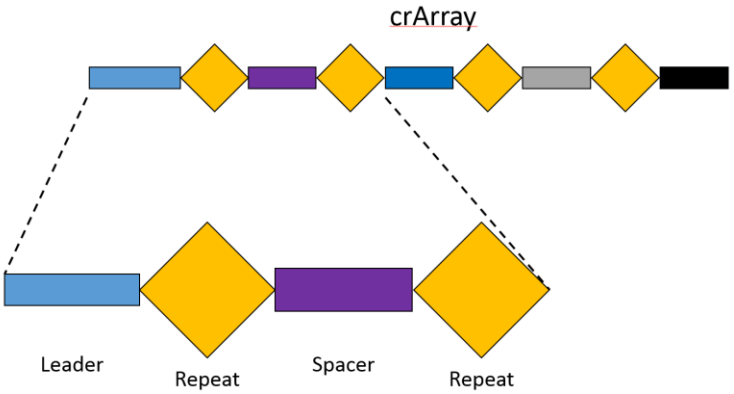
CRISPR systems mediate their function by processing the CRISPR array tandem transcripts into individual units, called CRISPR RNA (crRNA) (13). This processing is performed in various ways and by various RNAses, and likewise the repeat-derived region of the crRNA can vary from a structured element, as in most Class I Systems, to a simple sequence that forms a duplex with a Trans-Activating CrRNA (tracrRNA), as in some Class II (14, 15) (Figure 1.1b). The tracrRNA for some Class II systems serves as a means of validating that minimally-structured Class II crRNAs are properly loaded by verifying through base-pairing that the RNAs originated from the crArray and not some other improper origin.

CRISPR diversity is vast and the mechanisms employed are as well (16-19). CRISPR systems have been observed to have anywhere from one single effector protein to large multi-sub-unit complexes. These proteins associate with the crRNA and use the crRNA as guides to detect foreign nucleic acids. The proteins associated with CRISPR Arrays are called Cas (CRISPR-Associated) proteins (20). In all CRISPR systems, acquisition machinery is lumped together with the recognition and nuclease machinery in a genetic element called the *Cas* Operon (17, 20) (Figure 1.1c). Acquisition proteins are responsible for acquiring new spacers and specifically integrating them into the host genome at the CRISPR array (21-23).

The use of bacterial immune systems for technological applications has a robust history and has been integral to the advancement of biological science, starting with the first applications of restriction enzymes to map genomes in 1971 and progressing through the currently developing CRISPR/Cas

revolution (24). Surely, the full potential of exogenously-applied bacterial immunity mechanisms has not yet been depleted.

a)



b)

c)

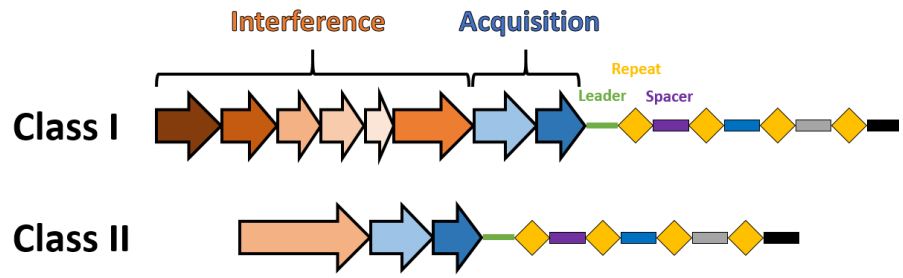


Figure 1.1: **Structure of CRISPR Genetic Elements. a)** The generalized composition which is shared by all crArray loci. The locus is comprised of a unique Leader sequence (blue) which is crucial for integration of new spacers into the array. The repeat sequence (orange) immediately follows the Leader and is also between each invader-derived spacer sequence (various colors). **b)** The Repeat region of the crArray in the Type I system, when transcribed into RNA, folds into a structured region. This stem loop serves as a recognition element for both the processing of the CrArray transcript into individual crRNA and as the first point of contact between the protein sub-units of the Cascade complex and the crRNA. **c)** The construction of the genetic elements of the CRISPR locus. The operon can be divided into groups of genes involved in either CRISPR interference or CRISPR acquisition. In the Class I CRISPR system, the interference proteins form a multi-subunit complex. In Class II, the interference modules are comprised of a single protein. In both classes, the Acquisition modules at the least contain the signature Cas1/Cas2 proteins, but may also contain others.

1.2 CRISPR Classification

One of the minor but seemingly constant hazards inherent to working in a new field as it matures is the evolving and conflicting terminology, classifications, and nomenclature. In this, CRISPR systems are no exception and, as an aid to the reader in following past literature, a brief review of the history of CRISPR and a summary of classification conventions is helpful. I will not endeavor to make this the most complete or nuanced version of such a review as others have recorded the history of CRISPR first-hand (25-28), and others have more exhaustively described the organization of CRISPR systems into their various Classes, Types, and Sub-Types (16, 17).

In short, and for quick reference, a table of the CRISPR systems most likely to be familiar to the reader is provided (Table 1). This does not represent the sum total of all CRISPR systems nor all Cas proteins, as new CRISPR systems are still currently being discovered. In some earlier literature, the seemingly deprecated term RAMP (Repeat Associated Mysterious Protein) was used to describe some Cas genes (29).

CRISPR systems are first divided into “Class”, of which there are only two recognized categories: Class 1 and Class 2. The Class 1 systems are those which have multi-sub-unit effector complexes and Class 2 systems have single-component effector complexes. Both require crRNA and rely on the same general principles of target recognition. The next-finer division occurs at the “Type”. There are 5 categorized types which are determined by the Class as well as the substrate (DNA/RNA/Both) and characteristic proteins of the CRISPR system (16, 17, 19, 30).

Table 1.1: **CRISPR/Cas nomenclature conventions.** Reference for the organization of CRISPR systems and for deprecated/less used names that appear in the literature.

Class	CRISPR Type	Target	Cas Protein Name	Other Names
1	I	DNA	Cse1	CasA, Cas8
			Cse2	CasB
			Cas6e	CasE, Cas6e
			Cas7	CasC, Cse4
			Cas5e	CasD
			Cas3[1]	
Shared Acquisition	-	-	Cas1	
Shared Acquisition	-	-	Cas2	
2	II	DNA	Cas9	
1	III	RNA[2]	Cas10	
2	V		Cas12a	Cpf1
	VI		Cas13b	C2C2

[1] Cas3 is, in some systems, split into Cas3' and Cas3'', splitting the protein into its helicase and nuclease domains

[2] Type III systems are also capable of cleaving either DNA, RNA, or both.

In this thesis, I will endeavor to maintain a continuous and consistent use of the naming conventions, with the primary convention of the Type I-E system being Cse1, Cse2, Cas7, Cas5e, and Cas6e. I will use a combination of *T. fusca* Cascade and *E. coli* Cascade as the representative systems of the Class I Type I-E system as they are the best characterized. I will use *S. pyogenes* Cas9 as the representative of a generic Class II Type II system for the same reason. Generally speaking, the Class 2 Type II system is the best characterized. This can be mostly explained by the fact that Cas9 genome editing technologies have necessitated and propelled advancements in the mechanistic understanding of Cas9. Also, since Class 2 Type II systems have a single-component effector complex, it is at first approximation easier to work with.

The first mention in the literature of what would later be called the CRISPR locus was in 1987, described as "An unusual structure… found in the 3'-end flanking region of *iap*… Five highly homologous sequences of 29 Nucleotides were arranged as direct repeats with 32 nucleotides as spacing." (31). At the early stages of CRISPR research, the first observation made which indicated that the CRISPR system might be some form of an immune system was the identification of the spacer sequences as being viral genome-derived (8, 32, 33). The CRISPR locus had already been identified as a repetitive region of DNA that was likely to be structured. Strains of bacteria which contained CRISPR systems were resistant to infection by phages or plasmids when the CRISPR region contained spacers originating from the invader, all but confirming the role of the CRISPR locus as an immune system (8). Further genetic evaluation showed that the associated *Cas* genes were necessary for the immune system to function (34). The determination that the *Cas* genes were essentially RNA-programmable nucleases led to the application of Cas9 for gene editing by both the Doudna and Zhang labs (35, 36). The resulting patent dispute is well documented in the literature as well as in popular science reporting (37-41). CRISPR has even penetrated popular culture with varying degrees of accuracy in the representation of CRISPR-based technologies – from CRISPR's dystopian influence depicted in speculative fiction novels such as *Change*

Agent by Daniel Suarez, or as the unifying theme of a crime-drama executive produced by Jennifer Lopez, to using CRISPR to create giant city-destroying monsters in the 2018 major motion picture *Rampage* starring Dwayne "The Rock" Johnson (42-44).

1.3 CRISPR Mechanisms

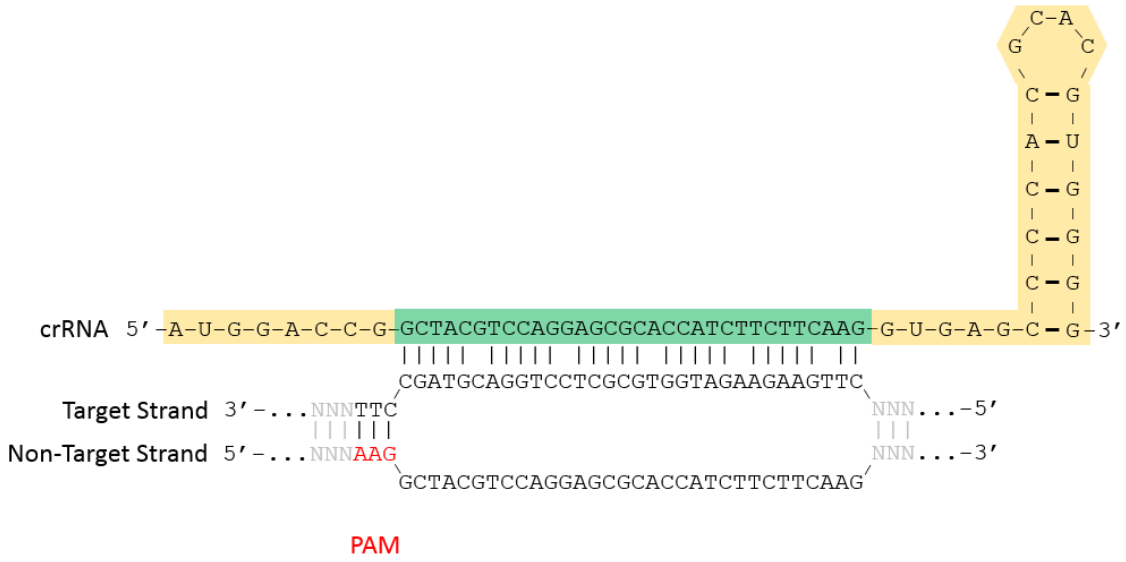
The absolutely conserved mechanism characteristic of CRISPR systems is RNA-mediated target recognition. Without RNA guides, CRISPR systems cease to function since these systems rely on RNA to provide specificity, flexibility, and affinity for various target invader nucleic acids. These mechanisms can provide interference against invaders possessing either an RNA or DNA genome. Another key feature that is strongly conserved is the Protospacer Adjacent Motif (PAM). This small sequence adjacent to the target site is required for target recognition in DNA-targeting CRISPR systems (45, 46).

The central mechanism of target recognition is direct base-pairing between the guide sequence and the target (7). This process is accomplished by progressive unwinding of the substrate DNA and substitution of base-pairing with the RNA guide – this is thermodynamically favorable as RNA-DNA heteroduplex is stronger than DNA-DNA (47, 48). Since the RNA guide itself is also strongly held by the Cas machinery, generally through non-specific electrostatic interactions, mostly to the phosphate backbone, the base pairing between the guide RNA and the substrate results in a strong association between the complex and the substrate nucleic acid (49, 50). In this, too, variations exist in CRISPR systems in terms of the length of the guide, degree of base-pairing, and the base-pair fidelity necessary to engage a full R-loop (51-54). For instance, the spCas9 guide is 22 nucleotides long and base pairing is uninterrupted throughout the length (49). In contrast, the Type I systems have variable target lengths. The *tfu*Cascade and *eco*Cascade guides are 32 nucleotides long, but every 6th nucleotide is flipped away from the target sequence (50, 55-58). This results in a region of specificity that is 27 bases long. The Type I-C CRISPR system of *Bacillus halodurans* has a variable spacer length, ranging from 32-36 nt (53, 59). It appears to

be a general principle of CRISPR target recognition that base-pair mismatches between the target and crRNA on the PAM-distal side is less detrimental than those at the PAM-proximal side (60, 61). This is because R-loop formation, and by proxy target recognition, occurs directionally, starting at PAM recognition, formation of a short region of critical base-pairs between the crRNA and the target called the seed-bubble, and propagating down the length of the recognized region (62). Mismatches in the seed bubble region carry a heavy penalty for target recognition and mismatches lead to an increased chance of rejection of the substrate (58, 60, 62, 63).

Throughout this work and in much of the literature, targeted DNA sequences are defined in terms of the Target Strand and Non-Target Strand (Figure 1.2a). The Target Strand is that which is directly base-paired to the crRNA. The Non-Target Strand is that which has been left as a single-stranded region. The Target Strand/Non-Target Strand convention has no bearing on which strand is modified by the CRISPR/cas system or related technology, and is a common source of confusion when discussions of CRISPR mechanisms are conducted with experts from outside the field. PAM Sequences can be located on either side of a targeted site depending on the CRISPR system, and are defined by the sequence on the non-target strand (49, 55, 64, 65).

a)



b)

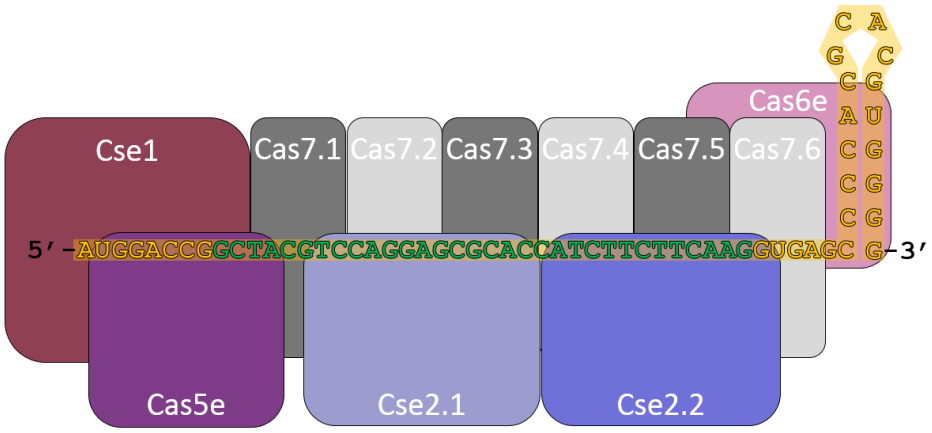


Figure 1.2: **Cascade target recognition and structural schematic. a)** Schematic of target recognition in the Type I system. The substrate DNA is defined in terms of the Target (the strand which participates in base-pairing to the crRNA) and Non-target strands (the strand which is looped-out). The PAM sequence is defined on the Non-Target strand. In the Type I-E system, the crRNA base-pairing is discontinuous with every 6th nucleotide flipped away from the substrate. **b)** Schematic representation of the Type I-E Cascade complex. The type I-E Cascade complex in molar equivalents is comprised of $(Cse1)_1$, $(Cse2)_2$, $(Cas7)_6$, $(Cas5e)_1$, $(Cas6e)_1$. The Cas6e sub-unit which recognizes the stem-loop of the crRNA cleaves the crArray transcript into individual sub-units and remains associated. The rest of the Cascade complex is then assembled along the Cas7 backbone with an underbelly of Cse2. The complex is then capped by Cas5e and the last sub-unit to associate is Cse1.

PAM sequence requirements vary in terms of their placement relative to the target sequence, in terms of their sequence content, and in terms of their length. The well-characterized *S. pyogenes* Cas9 possesses a three nucleotide PAM sequence on the 3' end of the targeted sequence on the Non-Target Strand (66). Type I systems have their PAM sequences on the 5' end of the targeted sequence on the Non-Target strand (50, 55, 67). PAM Specificity and sequence content vary from system to system. For instance, the spCas9 canonical PAM is the NGG PAM (66, 68). The *tfu*Cascade canonical PAM is AAG (54, 69). The mechanisms of PAM recognition can also diverge between different systems. SpCas9's PAM recognition is mediated through contacts in the major groove between the PAM nucleotides and a conserved set of Arginines in the C-terminal domain of Cas9 (70). By contrast, *eco*Cascade, and most likely all other Type I-E systems at least, recognizes its PAM through contacts in the minor groove mediated by contacts to a glycine-rich loop present on Cse1 (55). In general, failure to recognize a PAM sequence results in failure to recognize a target (48, 71).

How can the PAM sequence be rationalized as such a strict target recognition requirement? After all, target specificity is ensured by the crRNA. PAM recognition primarily serves as a protection against self-targeting. Any crRNA sequence that encodes a guide against a target is a sequence that also exists in the host genome at the CRISPR array. If no PAM sequence were required, all surveillance complexes would target the host genome at this locus. In fact, in CRISPR systems that have a PAM requirement, the PAMs with the lowest degree of target recognition are also those sequences which are present in the repeat region of the CRISPR array (72). In other words, the "worst" PAM for interference is that which is contained at the CRISPR locus. An additional supporting argument could be made that the presence of a PAM sequence also speeds target recognition, since the recognition mechanisms for PAM sequences do not require target unwinding and PAM sequences are short. This could prevent surveillance complexes from spending too much time trying to recognize improper target sequences.

1.4 Type I-E CRISPR Structural Overview

Since this thesis is primarily concerned with the Type I-E CRISPR system, it will be discussed the most in-depth structurally. The first Type I-associated structure solved with functional information was Cas6 in *Pyrococcus furiosus* (73). In *Pyrococcus furiosus*, one gene copy of Cas6 is used to process crRNAs from the Type I-A, Type I-G, and Type III-B systems (74, 75). This serves as a useful example demonstrating that host genomes can contain multiple CRISPR systems and that some components, such as the processing or adaptation machinery, might be shared between them. Cas6 is responsible for recognizing the nascent crRNA and cleaving the long transcript into individual crRNA units by recognizing and cleaving the repeat elements specifically (76). In Type I-E systems (in which Cas6-like proteins are called Cas6e), once crRNA processing is complete the mature crRNA remains bound to Cas6e and nucleates the formation of the Cascade complex (CRISPR-Associated Complex for Anti-viral Defense) (77).

The first structure of a full Cascade complex was the *E. coli* Cascade complex (51). It showed the overall architecture of the complex with one copy of Cse1, two copies of Cse2, six copies of Cas7, one copy each of Cas5e and Cas6e (Figure 1.2b). The Cas5e and Cse1 are at one end and Cas6e is at the other. Between them, a twisting backbone of Cas7 and an underbelly of Cse2 subunits coat the crRNA along the region that contains the targeting information. Following this, a structure of the *E. coli* Cascade was solved that showed Cascade bound to a single-stranded target, revealing that the substrate and crRNA have every sixth nucleotide flipped away from one another, not participating in base-pair mediated recognition (50). This base flipping is a structural requirement, as the DNA-RNA heteroduplex is significantly distorted from ideal B-form DNA.

Hochstrasser *et al.* 2014 solved the first Cascade-Cas3 structure; the Cryo-EM map showed the type I-E system from *E. coli* (78). The resolution was poor but clearly showed that the Cse1 sub-unit was the point of recruitment for Cas3. *Tfu*Cas3's crystal structure was solved the same year and associated

biochemistry shows that the linker-helix motif of Cas3 is important for the recruitment of Cas3 to Cascade (69). Hayes *et al.* showed a more complete picture of target binding, using a partially double-stranded target (55). It revealed the PAM recognition elements as well as defined a conformational change in the Cse1 sub-unit C-terminal domain that would prove significant in Cas3 binding and validation of target recognition. Another set of cryo-EM structures using *T. fusca* Cascade showed partial- and full-R-loop formation on a full-length double stranded substrate (58). It shows conformational change mechanisms that signal Cas3 recruitment in more detail as well as hint towards a mechanism that helps Cas3 initial cleavage by bulging the non-target strand towards the expected site of Cas3's nuclease center. It also identified a series of salt bridges between Cse2 and Cas7 that are not present in *E. coli*. These salt bridges close over the target strand-crRNA duplex, locking it in place, but these interactions also increase the thermal requirement for target binding. Xiao *et al.* 2018 describes in more detail the Cas3/Cascade interface surfaces and provide residue-specific interaction mechanisms (57). These structures also show that the initial nicking of the substrate DNA occurs through direct interaction with the nuclease center, and the nicked substrate is only threaded through the helicase afterwards.

Cas3 itself is the nuclease that degrades DNA in the Type I-E system (69, 79, 80). It is comprised of a nuclease-helicase fusion and translocates processively along DNA in a 3' – 5' direction (Figure 1.3). The helicase is a SF2 family helicase with two RecA domains that work cooperatively to move along the substrate DNA. This translocation happens in three nucleotide bursts in *E. coli* (81). Other single molecule work has shown that Cas3 seems to have two modes of activity (82, 83) (Figure 1.4). In the first, it is recruited to Cascade specifically, making the initial nick and reeling the substrate DNA. This loops out the target strand since Cas3 is still bound to Cascade which tethers the supercomplex to the target site. The non-target strand is intermittently cleaved and some sections re-anneal after passing through the helicase, generating sections of double- and single-stranded DNA. In the second mode,

Cas3 has released from Cascade and translocates independently. *Tfu*Cas3 has been shown to have a highly processive helicase, capable of translocating over 10kb away from the Cascade targeting site (83). There have been some hints that the cleavage products of Cas3 have some form of sequence preference or might be involved in spacer biogenesis during CRISPR adaptation, though further studies are required (84, 85).

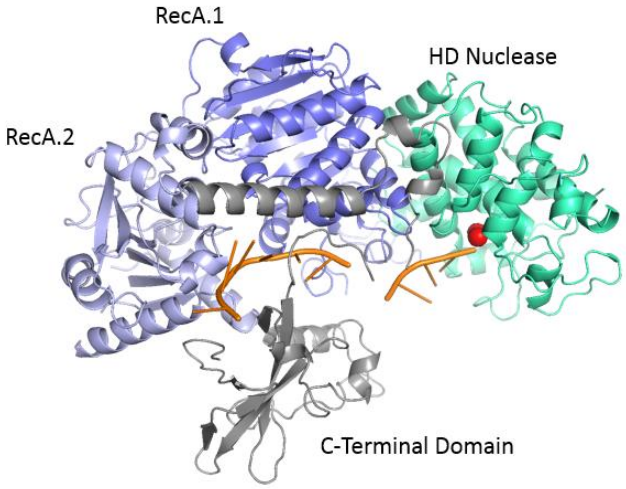


Figure 1.3: **Cas3 structure and mechanism of target degradation. a)** Domain organization of the Type I-E Cas3 nuclease-helicase. The HD-nuclease (teal) contains two divalent metal ions (red); the SF2 Helicase is comprised of two RecA domains (two shades of blue); and the Linker Helix and C-Terminal Domain (gray). The substrate DNA (orange) is threaded through an opening between the C-terminal domain and the helicase, along a substrate groove, and fed into the nuclease active site (adapted from the structure in Huo *et al.* 2014).

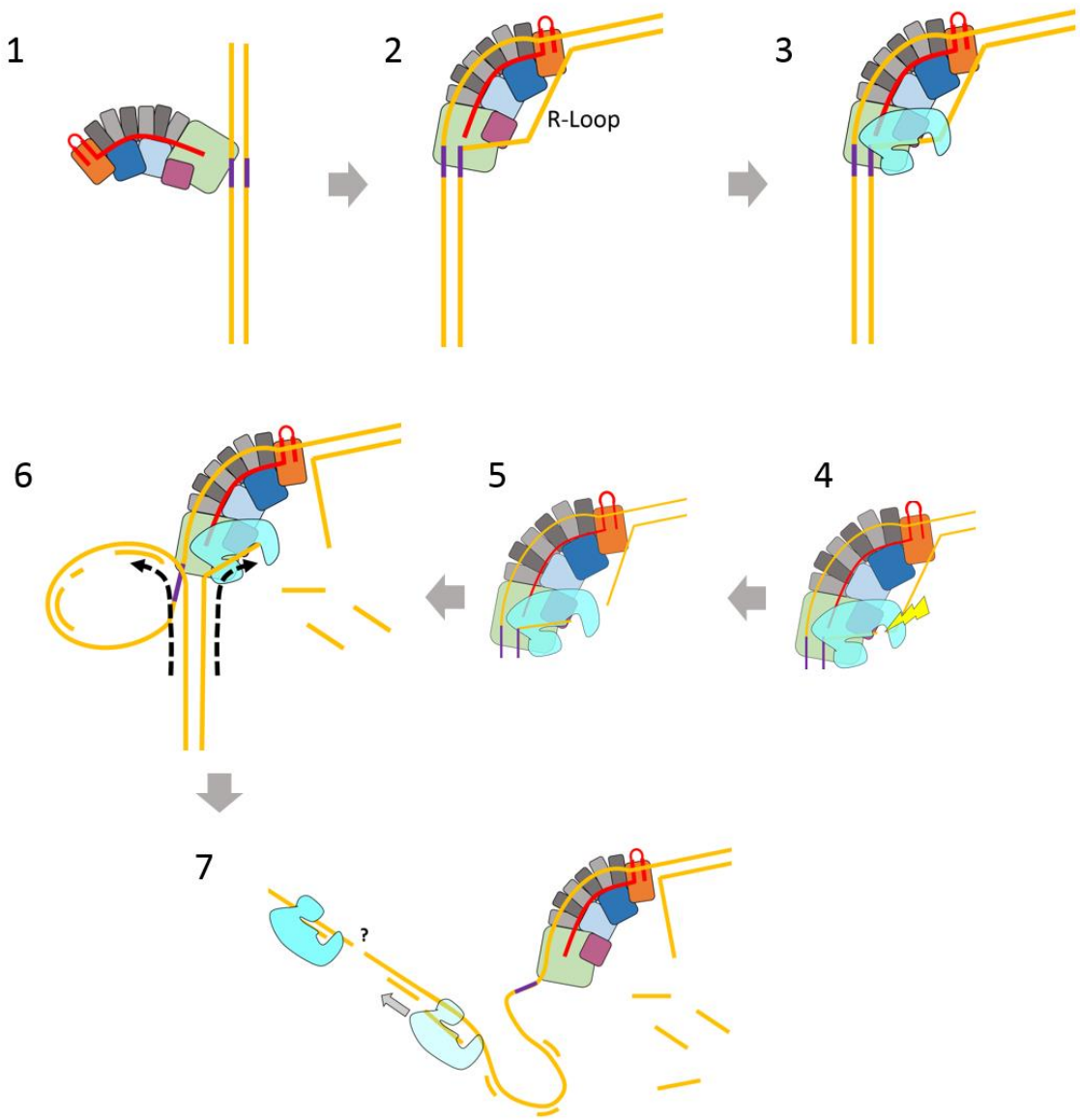


Figure 1.4: Steps of DNA interference in Type I-E CRISPR. 1) PAM recognition occurs mediated by contacts between the Cse1 sub-unit and the substrate DNA. If the PAM sequence (Purple) is not validated, the Cascade complex will dissociate from the substrate. **2)** If the PAM sequence is verified, direct base-pairing between the substrate and the crRNA form an R-loop. If base pairing is not validated to a sufficient degree, the R-loop collapses and Cascade dissociates from the substrate. Once proper recognition between the crRNA and the substrate occurs, a conformational change locks the substrate in place. **3)** Cas3 associates to the full formed R-loop, with the nuclease domain oriented towards the looped-out single-stranded DNA. **4)** A nick is produced on the looped-out DNA at approximately the 9th position from PAM. **5)** Cas3 loads the substrate through the helicase domain, probably by threading. **6)** In the first mode of activity, Cas3 remains associated to Cascade as it reels the DNA, producing a loop of DNA on the target strand which consists of a mixture of single and double-stranded DNA. **7)** In the second mode of activity, Cas3 dissociates from Cascade and translocates independently. It is possible that double strand breaks are formed at some points during this step.

Huo *et al.* revealed that the preferred divalent ion for the active site of *tfu*Cas3's HD nuclease domain was Co2+ for *in vitro* cleavage (69). When solving the crystal structure, the protein was recombinantly expressed in *E. coli* in LB media. The resulting structure shows two iron ions in the active site of the nuclease. Unfortunately, this iron-loaded *tfu*Cas3 is not very active and requires a very high concentration of Cobalt for in-vitro cleavage assays and requires high temperature cleavage conditions. This suggests that the metal ion at the active site is held very strongly by electrostatic interactions with the nuclease domain, making later replacement of the active site ions inefficient. To get around this I expressed *tfu*Cas3 in M9 media lacking a metal supplement, adding Cobalt at the induction of expression (Chapter III). This resulted in a drastic increase in activity, even when cobalt was left out of the in-vitro cleavage buffer. However, it is not conclusive that cobalt is the native metal ion for *T. fusca* cas3 in the host bacteria. Instead it is possible that Fe2+ is indeed the native ion, but in recombinant conditions iron is quickly oxidized to Fe3+[3].

1.5 CRISPR Adaptation

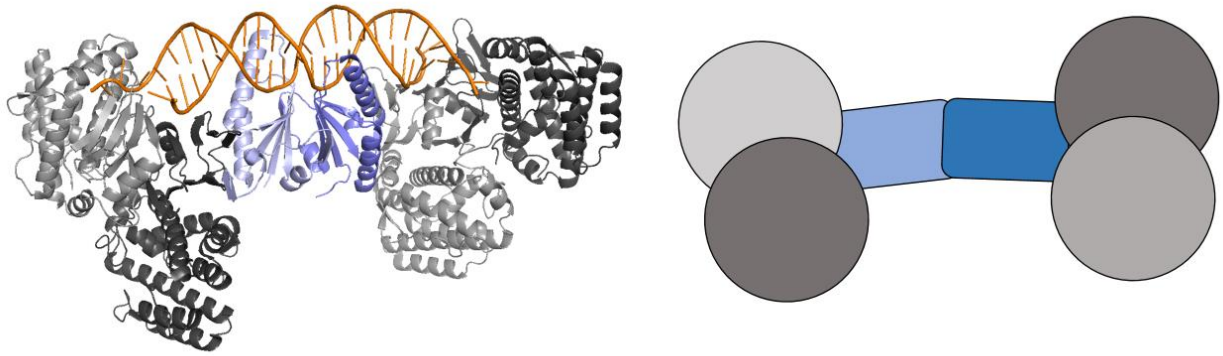
As an adaptive immune system, CRISPR arrays must contain a mechanism to update the library of targeted sequences in response to invasion. This process is not as well understood as CRISPR interference, with several outstanding questions such as the mechanisms of spacer biogenesis, primed adaptation, and the details of spacer insertion into the CRISPR array.

Most, if not all, CRISPR systems rely on a conserved pair of proteins, Cas1 and Cas2, as the core of their adaptation machinery (86, 87). These proteins form a complex with 4 copies of Cas1 and two copies of Cas2 (Figure 1.4a). The arrangement of this complex is in a "dumbbell", with two Cas1 dimers bridged by a dimer of Cas2 (21). Cas1 contains a nuclease site, giving the overall complex a total of four nucleases (88). The size of the complex acts as a "ruler" to measure the length of a spacer via non-

[3] Personal communication.

specific contacts to the substrate DNA (89). Integration into the crArray occurs at the site of the first repeat, which is next to a sequence called the Leader (34) (Figure 1.4b). This Leader determines the orientation of insertion, since orientation is vital to proper CRISPR interference from the resulting crRNA (23). Upon integration, the first repeat is split into two complementary regions which remain base-paired but that bridge the leader to the spacer and to the rest of the crArray (90). In some systems, the Cas1-Cas2 complex is sufficient for integration into a crArray, host factors are required in others (23). Through a mechanism that is not fully understood yet, the repeat regions are repaired, resulting in copying of the repeat, integration is complete, and the crArray is expanded to include a new spacer (91). In some Type II CRISPR systems, an additional component is required for spacer acquisition but is seemingly not involved in integration, Csn2 (92, 93). This protein is a tetramer and has a putative function in spacer processing prior to integration. There is evidence to suggest that it bridges Cas1/Cas2 to Cas9 and serves to bring the whole complex together during acquisition and Cas9 likely plays a role in integration target selection (94).

a)



b)

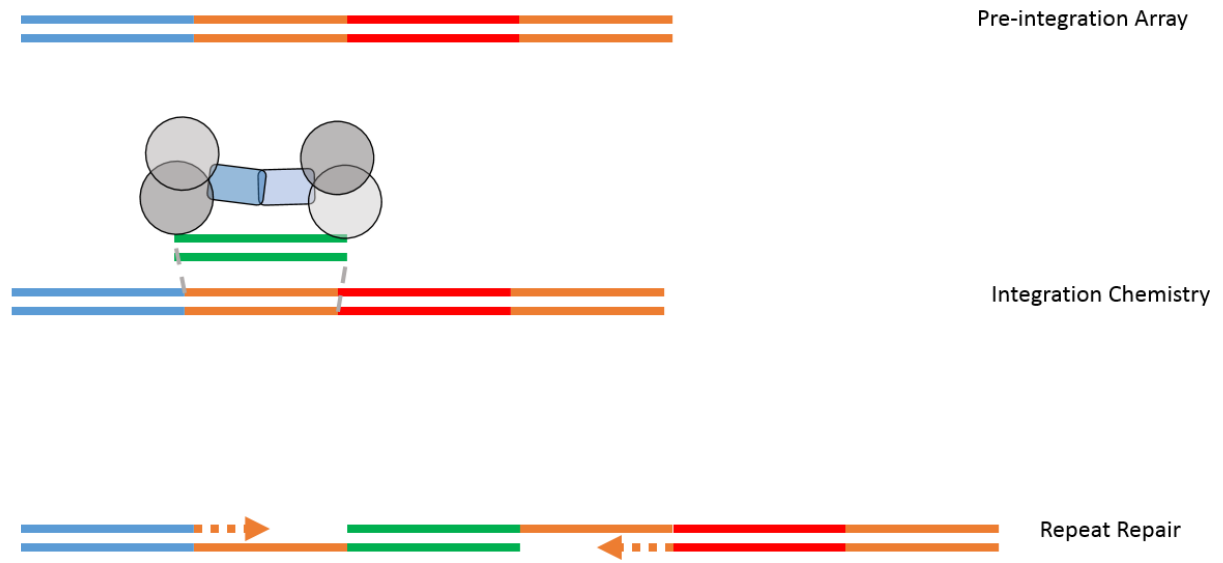


Figure 1.5: **Cas1/Cas2 organization and model of spacer integration. a)** Cas1-Cas2 structural organization shown as a crystal structure (left) and schematic (right). Cas1 sub-units are shown in Gray and Black, Cas2 subunits are shown in blue. Substrate spacer DNA is shown in orange. The complex is organized as two dimers of Cas1 bridged by a dimer of Cas2 (adapted from the crystal structure in Xiao *et al.* 2017). **b)** Schematic representation of integration intermediates during Cas1/Cas2-catalyzed spacer integration. Cas1/Cas2 recognizes the leader sequence and binds to the repeat immediately adjacent. Next, it catalyzes integration such that the new spacer is ligated to the repeat at the 5' ends, resulting in a duplication of the repeat sequence after an uncharacterized repair event.

Through a mechanism not fully understood, all spacers that are integrated into crArrays contain interference PAMs adjacent to their target site in the invader genome. This ensures that interference is possible against the target site. There is some speculation in the literature that Cas1/Cas2 contains a PAM recognition motif, but the evidence for this is not complete (89). There is, however, another mechanism that compensates for the possibility of escape mutations at target sites of invader genomes. This mechanism is called Primed Adaptation or Primed Spacer Acquisition. In this process, a partially mismatched target sequence or a mutated PAM sequence can trigger the acquisition of new spacers at a higher rate than naïve adaptation (84). This process is Cas1/Cas2, Cascade, and Cas3-dependent in the Type IE system, but the detailed mechanisms are not understood (95). In the Type I-E CRISPR system, RecBCD helicase activity is a requirement for naïve adaptation (96). RecBCD is a nuclease-helicase complex capable of moving in both 5'-3' and 3'-5' directions with two separate helicase domains (97). It has been speculated that Cas3 might be fulfilling a similar role to that of RecBCD in the selection of pre-spacer substrates for Cas1/Cas2 during primed adaptation (85).

1.6 CRISPR Applications

An intense focus of CRISPR research has been the development of new molecular biology technologies. As an indicator of this trend, many basic science CRISPR research articles tend to have passages in the discussion section which propose possible implications or technologies that might arise from the research. Most CRISPR-based technologies rely heavily on the principle of RNA targeting and how easy it is using CRISPR/Cas to target unique and varied sequences with a protein effector, relative to other existing technologies. These technologies take the form of direct substrate cleavage or modification as well as co-localization and imaging. Nucleic acid detection has also been explored through CRISPR technologies that enable novel approaches with high sensitivity.

Genome Editing

Cas9 and other single-component effector complexes are natural gene editing tools. They are highly programmable in a way that other gene editing technologies have not been traditionally. For instance, in order to induce a double-strand break at a specific site, one could employ a restriction enzyme, of which there are many high-fidelity engineered nucleases available. However, due to the targeting size (typically 6-10bp), in a large genome such as a mammalian genome, there are inevitably redundant sites and it is unlikely that a restriction enzyme exists for your desired target site. Meganucleases are along the same vein and go partway to solving the specificity problem but again are limited in the sequences that they can target. Alternatively, using TALENs creates a much more highly specific system, but they require a lot of cloning steps to assemble the domains in the correct order to confer specificity.

In comparison, Cas9 is extraordinarily simple to use. An end-user of a CRISPR technology typically only needs to make minor considerations regarding their target site, such as the presence of a PAM nearby. And, given the small size of the PAM requirement, PAMs are not a rare occurrence.

Cas9 genome editing technologies have been applied to a wide range of targets, cell types, and organisms. The first editing studies were conducted in cancer cell lines by both Zhang (Cong *et al.*) and Doudna labs (Jinek *et al.*) both published in January 2013 (36, 98). Subsequently, mutant mouse generation quickly followed (99). Since then, Cas9 editing has been performed on primary cell lines, embryos, and in various model organisms from Drosophila to Zebrafish. Derivative technologies have been developed that utilize the CRISPR/Cas9 platform and accelerated the trajectory of progress in the biological sciences. These other technologies involve screening applications for multiplexed gene editing; active domain fusions to localize biochemical activity to a target site; engineered genetic elements called “gene drives” which can theoretically push evolution uphill by forcing the fixation of a

detrimental allele. All of this was possible from the development of a single editing platform. The availability of other platforms might help this process to develop even further.

Engineering Strategies

Modifications to the natural CRISPR/Cas machinery help to expand the capability or the ease of use for CRISPR/Cas-based technologies. These approaches fall under three main umbrellas: simplification of the system, improved fidelity, or expanded PAM recognition. In these, the simplification of the Cas9 machinery was accomplished by creating the single-guide RNA. This is a fusion of the tracrRNA with the crRNA at the end of the stem that mediates the tracrRNA/crRNA interaction in the natural case. This modification allows a single transcription unit to encode the entire Cas9 guide (66).

Further modifications directly to the Cas9 protein have been performed to increase the fidelity of the enzyme in its target recognition. This has generally been accomplished through mutations that slow down the conformational change during the target recognition step, causing the active conformation to be less favored during off-target recognition events (100, 101). This is important because conclusions have been murky when it comes to the off-target effects of Cas9 activity, with studies going back and forth, being published and retracted alternatively. As of yet, I would not say there is a well-settled conclusion of off-target activity for *in-vivo* CRISPR/Cas target recognition or a full characterization of the nature of unintended editing events (102-108).

In the pursuit of unique activity or in reducing off-target effects, there are many ways to achieve more favorable biochemistry. If a structure or enough structural information exists to perform rational design engineering, this can speed the process along greatly. In the most basic example, in using Cas9 as a targeting platform without any need of the nuclease activity, single point mutations at the nuclease reaction centers were performed (109-111). Along these same lines, deactivating only one nuclease site or the other can be used to generate a gentler nickase as opposed to a double-stranded endonuclease

(112-114). In more complicated scenarios though, there are many mutations or modifications that exist which might assist towards the goal but are not immediately obvious. For this reason, homologue screening and directed evolution are two methods which can provide a framework for success.

Homologue screening strategies have primarily been employed as a means to expand the library of PAM sequences that are targetable by CRISPR-based genome editing strategies. While PAM sequences are common in most genomes, the availability of a PAM site at a specific locus limits the precision of editing technologies. If an expanded library of CRISPR/Cas systems exist, it inherently expands the availability of target sites, since PAM sequences are not entirely conserved across diverse CRISPR systems. For instance, the *N. meningitides* Cas9 has a NNNNGATT PAM (115). The increased length of the *N. meningitides* Cas9 PAM decreases the frequency of a PAM site, but also theoretically decreases off-target effects, and is compatible with lower-GC content regions of the genome (116, 117). Additionally, the exploration of other CRISPR systems with different biology entirely can move the field forward. The more recently discovered Cas12 of the Type V systems is one such exploration, with its development still in the early stages (118). Instead of generating a blunt double-strand break, Cas12 (previously Cpf1) generates 5 nucleotide 5' overhangs after target validation, similar to more traditional restriction enzymes commonly used for cloning. Cas12 recognizes PAM through contacts on both the major and minor groove, but the PAM identity is very different (Cas9: NGG, Cas12: NTTT) and recognition occurs on the opposite side of the target site compared to Cas9 (64). Cas12 has a distinct potential advantage in that the crRNA does not require a tracrRNA and Cas12 is capable of processing its own crRNA whereas Cas9 requires other processing factors (119). However, when Cas12 binds to a cognate target substrate, the protein's nuclease behavior is not only localized to the target site, but instead acts seemingly indiscriminately on all nearby ssDNA substrates (120). This has largely halted the development of Cas12 as a gene editing platform.

Directed evolution strategies employ an artificial selection mechanism to achieve a desired biochemical activity. The basic principle is that a library of candidate constructs (generally produced through random mutagenesis) are generated and the relative abundance of the members of this library undergoes a selection scheme (121). For instance, in selecting for a more active protein, Hu *et al.* employed a design whereby a library member's improved ability to survive phage infection resulted in a higher representation in the population of host *E. coli* used to conduct the experiment (122). After selection, PAM specificity could be altered from NGG to include NG, GAA, or GAT.

Nucleic Acid Detection

The ability to reprogram CRISPR/Cas systems to target any arbitrary nucleic acid makes them attractive for nucleic acid detection platforms. These systems would be useful in diagnosis of viral diseases when viral loads are small or to differentiate between different but related viruses. Furthermore, some CRISPR/Cas systems, such as Cas13 of the Class 2 Type VI systems or Cas12 of Class 2 Type V, exhibit non-specific and rampant collateral trans-cleavage of other nucleic acids (120, 123). The generalized set of technologies involves Cas13 or a CRISPR system capable of indiscriminate cleavage upon target binding in solution with a sample of nucleic acids and a reporter fluorophore linked to a quencher. When Cas13's cleavage activity is activated by binding to a target, the fluorophore reporter is cleaved through non-specific activity and increases the fluorescent signal in solution. This approach is capable of approaching atto-molar sensitivity since a single activated Cas13 protein can cleave many fluorophore reporters and thereby amplify the signal to the detection threshold (124, 125). Similar technologies have been developed based on Cas12 (125-127).

Domain Fusions

Due to CRISPR/Cas technologies being so flexible in their programmable targeting, they are attractive scaffolds to be used in delivering protein domain payloads with biochemical activity to defined nucleic

acid sequences with a high degree of specificity. This can be accomplished by designing domain fusions of enzymes or other epitopes that confer novel functions to the Cas machinery. Most often, this is done with catalytically dead Cas9 (dCas9).

Epigenetic modifiers or transcriptional regulators are a common domain fusion (128-130). One of the first non-genome editing approaches developed for the dCas9 platform is the technique known as CRISPR Interference or CRISPRi (111, 131). This CRISPR interference is separate and apart from the interference discussed in general CRISPR mechanisms, wherein DNA degradation is the mechanism of interference, though they share the same name. Instead, the term is meant to draw an analogy to the genetic tool of RNAi, wherein exogenous short RNAs are used to mediate mRNA cleavage and thus expression repression. In CRISPRi, transcriptional repressor or activator domains are fused to dCas9, thus allowing artificial transcriptional control at a target promoter. This can be used to increase or decrease the expression of a target gene to observe how modulation effects cellular processes, or it can be used to artificially bring a transcription factor to a promoter of a reporter gene to measure the effect a particular factor has on transcription.

An application that is much needed in the field of genome editing is the ability to precisely and specifically modify a base to a desired identity. To address this, cytidine de-aminase has been fused to Cas9, conferring the ability to de-aminate substrate cytidine, resulting in a transition to a uridine base (132-135). This uridine is then repaired to a thymine by repair mechanisms which excise and replace DNA uridine with thymine (136). Since this modification can occur on either strand, G->A modification is also possible and thus it allows a not insignificant range of editing targets. This activity is not as yet specific enough to modify a single base and only a single base, but it highlights how much closer the engineered tools are to solving these problems.

Precision Medicine

One of the most anticipated applications of precision gene editing has been the modification of a patient's own stem cells or the editing of patient-derived induced pluripotent stem cells. This application has the potential to treat a slew of genetic disorders which affect a subset of tissues due to a mutation, such as many blood diseases (hemophilia, sickle-cell, etc.) (137, 138). Individual genetic surgeries can repair the mutant gene to an active form in cell culture and the edited stem cells can then be returned to a patient by transplant (139, 140). There, the edited cells will differentiate, creating healthy tissue in which the deficiency has been corrected. Using a patient's own stem cells or patient-derived induced pluripotent stem cells reduces the risk of graft rejection and greatly reduces the risks of therapy (141). This approach has seen promising results in rescuing even otherwise lethal mouse models of disease and phenotype corrections have been observed in cell-culture of patient-derived primary cell lines (142-147). Clinical trials of these procedures are upcoming and present an exciting synthesis of both induced pluripotent stem cell technologies and CRISPR gene editing.

Cas9 for Large Deletions

Using Cas9 for deletions of gene targets are usually accomplished through introducing point mutations at the target site via repair processes that introduce frame-shifts to the coding sequence (148). This is useful for many applications from developing disease models, functional genomics investigation, and even the potential rescue of dominant-negative disease phenotypes. However, there are a number of applications where frame-shift mutations or other small mutations are not sufficient to cause the desired effect, such as when investigating regulatory sequences, enhancers, or tandem repeat regions of the genome. For these applications, one needs to fully delete the whole genomic locus that comprises the element of interest. This can be done with Cas9 approaches, such as dual-targeting with two Cas9 guides. This however, has a low efficiency compared to standard editing, and the larger the region to be deleted, the lower the efficiency is (149-151). This has consequences in modeling cancer genetics, since many cancer phenotypes are caused by large genomic deletions (152-154).

Type I CRISPR Technologies

In this thesis, I describe a novel technology based on the Type I-E CRISPR system of *T. fusca*. Though it is the first technology to utilize the type I system outside of bacteria to which those systems are endogenous, it is not the first technology or technique developed.

The earliest technologies to utilize the Type I CRISPR system were as a CRISPRi platform. CRISPRi is a technology originally developed using a catalytically inactive Cas9 which utilizes a guide sequence against either a coding region of DNA or a promoter element to cause a repression of transcription in that region (131). This is achieved either through direct obstruction of the DNA by the bound CRISPR-cas complex or by a fused repressor domain. Since this technology does not require anything more than the Cas proteins and a provided targeting crRNA expression cassette, it is possible to re-purpose existing CRISPR systems in endogenous host bacteria for CRISPRi strategies. This was performed by both Luo *et al.* 2015 and Rath *et al.* 2015, where both used the *E. coli* Type I-E system (155, 156). Both approaches utilized a Cas3-knockout strain so that target recognition by the endogenous Cascade complex would not result in target cleavage. In both cases, the authors introduced the crRNA through plasmid transformation and observed repression whether Cascade targeted the coding region or the promoter of a reporter gene. Both papers observed that the effectiveness of repression depended on the strand that Cascade was bound to, but their results were conflicting. Rath *et al.* 2015 found that targeting the template strand of their GFP reporter resulted in more repression (155). Luo *et al.* 2015 found that targeting the non-template strand resulted in stronger repression, which is the same observation that has been made for dCas9-based CRISPRi technologies (156). Overall, since Type I CRISPR-system containing bacteria represent a majority of those bacteria with any CRISPR system at all, the ability to use the native Type I system as an investigative tool expands the capabilities of researchers who work in non-model organisms (30).

The only other genome engineering technology that exists using Type I CRISPR systems is that which was developed by Li *et al.* 2016 (157). They developed a method that uses the CRISPR systems native to *Sulfolobus islandicus* to enforce genetic selection for desired knock-in or knock-out mutations. Their design relied on the delivery of a plasmid which contained a donor DNA sequence with homology arms for the mutation target site and a minimal crArray which targets the un-mutated genomic site. The crArray in the plasmid gets transcribed and is then loaded to either the Cmr or Cascade complex. If homology directed repair did not occur, the CRISPR system binds the genomic locus, causing chromosomal degradation and destruction of the host genome. This means that the vast majority of surviving cells are the result of homology directed repair. In this way, the CRISPR system can be said to be used more as a mechanism for negative selection rather than as a true genome editing tool.

In this thesis, I describe the first true genome editing technology that uses any Type-I CRISPR system. In human embryonic stem cells, the use of Cascade and Cas3 generates large genomic deletions with variable start sites, variable end sites, and variable sizes. Cascade is tagged with Nuclear Localization Signals (NLS) on the C-terminal end of Cas7, Cas3 is also tagged with NLS on the C-terminal end. Cascade is co-expressed along with crRNA guides that target a genomic locus. When delivered via electroporation, the system introduces directional deletions in the expected direction of Cas3 translocation. These deletions seem to often be scar-less, with no inserted nucleotides resulting from the repair mechanism detected in the deletions that were analyzed by sequencing. While the start, stop, and length of deletion cannot be controlled at this time, this application allows for the creation of large genomic deletions with a single guide RNA. In *E. coli* and *T. fusca* Type I systems, both AAT and AAG are interference PAMs due to the promiscuous nature of Cascade PAM recognition (158). These PAMs are both compatible with low-GC regions of chromosomes, unlike the NGG Cas9 PAM. This could be relevant because the human genome has 100kb stretches that range in GC content wildly with an average of 41% GC (159). Even though interference for AAT is measured as lower than AAG for *T. fusca*

in an interference assay, Cas9 is unable to target any site that lacks a GC pair – even with extensive engineering to alter PAM specificities (69, 160).

In the following chapters of my thesis, I will discuss three lines of investigation. In the first I discuss mutational analysis of the Cas3-Cascade interface and how it suggests a complex mode of interaction during the recruitment of Cas3. In the second, I describe a genome editing technique we developed that uses the Type I-E CRISPR system from *T. fusca* to generate long genome deletions in eukaryotic cells. In the third, I outline a model to describe the mechanism of Cas3-induced double strand breaks via dimerization of Cas3 independent of Cascade.

CHAPTER II CREDITS

CREDITS:

This chapter describes perturbation of the interface between Cas3 and Cascade and the effects on target binding and cleavage which resulted. It is adapted from an unpublished manuscript. As described in the author contributions section, Yibei Xiao, Ailong Ke, and I designed the experiments and contributed to the interpretation of the data. I performed the experiments and wrote the manuscript.

CHAPTER II

Interface residue perturbation suggests a complex interaction mode in Type I-E Cascade-Cas3 complex

Adam Dolan[1], Yibei Xiao[1], and Ailong Ke*[1]

[1] Department of Biochemistry Molecular and Cell Biology, Cornell University, Ithaca, New York, 14850, USA

* To whom correspondence should be addressed. Tel : 1+ (607)-255-3945 ; Fax : 607-255-6249 ; Email : ak425@cornell.edu

Present Address: Ailong Ke, Department of Biochemistry Molecular and Cell Biology, Cornell University, Ithaca, New York, 14850, USA

2.1 ABSTRACT

CRISPR (clustered regularly interspaced short palindromic repeats) loci and the nearby CRISPR-associated *(cas)* operon provide bacteria and archaea with an RNA-based immunity system against foreign genetic elements. In Type I CRISPR-Cas systems, the crRNA-containing Cascade complex generates an R-loop structure in the matching region of a foreign dsDNA, and subsequently recruits the nuclease-helicase fusion enzyme Cas3 for processive DNA degradation. Despite multiple efforts, a clear definition of the Cascade and Cas3 interaction remained at low-resolution. Here I show that site-directed mutagenesis around the Linker helix of Cas3 (residues 779-797) perturbed its binding with Cascade. Whereas some mutants led to losses in both Cas3-binding and DNA target cleavage, others enhanced Cas3-binding, but abolished DNA cleavage. These different behaviors suggest an unexpected complex mode of Cascade-Cas3 interaction. I speculate that in addition to the direct physical interaction

mediated by the Linker helix of Cas3, the conformational state in the helicase of Cas3, and the relative orientation between Cas3 and Cascade, may play important roles in Cas3 recruitment and activation.

2.2 INTRODUCTION

Nearly 50% of bacteria and 80% of archaea organisms maintain the Clustered Regularly Interspaced Short Palindromic Repeats (CRISPR) and the associated *cas* operon nearby to provide an RNA-based adaptive immune system against foreign nucleic acids in bacteriophages and conjugating plasmids **(161-163)**. CRISPR associated proteins (Cas proteins) utilize the processed CRISPR RNA (crRNA) as a guide to mediate surveillance for, interference against, and acquisition of foreign nucleic acid molecules **(161-164)**. CRISPR systems are further classified into types based on the signature proteins in their *cas* operons; each type uses a distinct interference mechanism **(165)**. Type I CRISPR system accounts for 90% of all naturally occurring CRISPR systems. It contains the signature protein Cas3, and can be further categorized into six subtypes based on distinctive features in the *cas* operon (Type IA – IF) **(165)**. Organisms such as *E. coli* and *T. fusca* utilize the Type I-E CRISPR system, which employs a CRISPR-associated Complex for Anti-viral Defense (Cascade) to search for dsDNA targets matching the spacer region of the CRISPR RNA (crRNA) **(51, 166, 167)**. The Type I-E Cascade consists of a crRNA guide and eleven protein subunits from five Cas proteins (Cse1 (1), Cse2 (2), Cas5e (1), Cas6e (1), and Cas7(6), and CRISPR-derived RNA) **(168, 169)**. Upon encountering a matching dsDNA adjacent to an optimal PAM (Proto-spacer Adjacent Motif) sequence, Cascade initiates the DNA unwinding process, promoting segmented base-pair formation between the crRNA spacer and the target DNA strand **(55)**. This loops out the non-target DNA strand, giving rise to an R-loop structure **(78)**. The Cascade-marked R-loop structure recruits the signature protein Cas3, which is a fusion of an HD nuclease and a super family II helicase **(69)**. Upon binding, Cas3 nicks the non-target DNA strand ~7-12 nucleotide downstream of PAM and processively degrades the non-target strand and target strand DNA in a sequential fashion **(170)**.

The sequential activation of target searching and degradation machines in Type I CRISPR systems may provide higher selectivity to the system.

How does Cas3 selectively bind to the R-loop forming Cascade, while avoiding the free Cascade? The mechanism remained poorly understood despite intensive biochemical dissection and structure-function analysis. In a low-resolution reconstruction of the Cascade-Cas3 complex, residual Cas3 densities were found on top of the Cse1 subunit of the Cascade **(78)**. The resolution of this EM reconstruction precluded detailed dissection of molecular interactions. A high-resolution crystal structure of the *T. fusca* Cas3 has been determined, and its interaction with the *T. fusca* Cascade reconstituted **(69)**. In subsequent mutagenesis experiments, it was shown that deletion of the linker helix in Cas3 (residues 779-797) resulted in a complete loss of Cascade-Cas3 interaction *in vitro*, and a loss of CRISPR interference function *in vivo* **(69)**.

To further dissect the Cascade-Cas3 interaction, we performed manual docking of the *T. fusca* Cas3 crystal structure into the *E. coli* Cascade structure bound to a partial R-loop **(55)**. Based on this model, a series of mutations along the Cas3 linker helix were designed and their effects on Cascade binding and DNA target cleavage were examined. Mutants were assessed for Cascade binding via EMSA and Cascade-mediated DNA cleavage activity via denaturing PAGE. We identified multiple mutants which affected defects in both assays and one mutant which surprisingly showed an increase in Cascade binding but abolished DNA cleavage. We were able to narrow down one point mutant in a region just N-terminal to the Linker Helix on the surface between the two Helicase domains which demonstrated a Cascade affinity defect.

2.3 RESULTS

Besides our understanding that Cas3 binds to the vicinity of Cse1 subunit of Cascade, little was known about the detailed molecular interaction. Our structure-function analysis of Cas3 identified its linker

helix as a key structural element mediating the Cascade-Cas3 interaction **(Figure 2.1a) (171)**. We therefore attempted a few manual docking exercises, assuming the linker helix is a major element at the interface. The Cascade structure model we generated is primarily composed of the *E. coli* partial R-loop crystal structure **(55)**. We then replaced the Cse1 component of this structure with a homology modelled *T. fusca* Cse1 **(172)**.

This model makes several assumptions that were supported by the known mechanisms of Cas3-mediated target degradation as well as by density shown in the published low-resolution Cas3-Cascade EM map (78). First, the orientation of Cas3 must agree with its direction of movement on the non-target DNA strand, hence the helicase is facing the PAM-proximal region. Second, the HD nuclease of Cas3 must be able to reach the non-target strand DNA 9-12 nucleotides downstream of PAM, assuming this DNA strand remains flexible in the R-loop region **(78, 170)**. Third, because Cas3 only binds to the R-loop presenting Cascade, we assume the Cas3-binding interface must span an area in Cse1 that undergoes conformational changes during the R-loop formation process **(55)**. These assumptions led to a manually docked model where Cas3 docks its Linker helix into a shallow groove in Cse1. This manually docked model agrees closely with the later-determined high resolution cryo-EM structure of Cas3 bound to Cascade **(Figure 2.1b) (57)**. The model was not generated in a systematic fashion, and it requires the non-target DNA strand to deviate from the true path slightly to access the Cas3 HD nuclease center. It was nonetheless very helpful in guiding the focus of the regions to target and our initial design of Cas3 mutants **(Figure 2.1c)**.

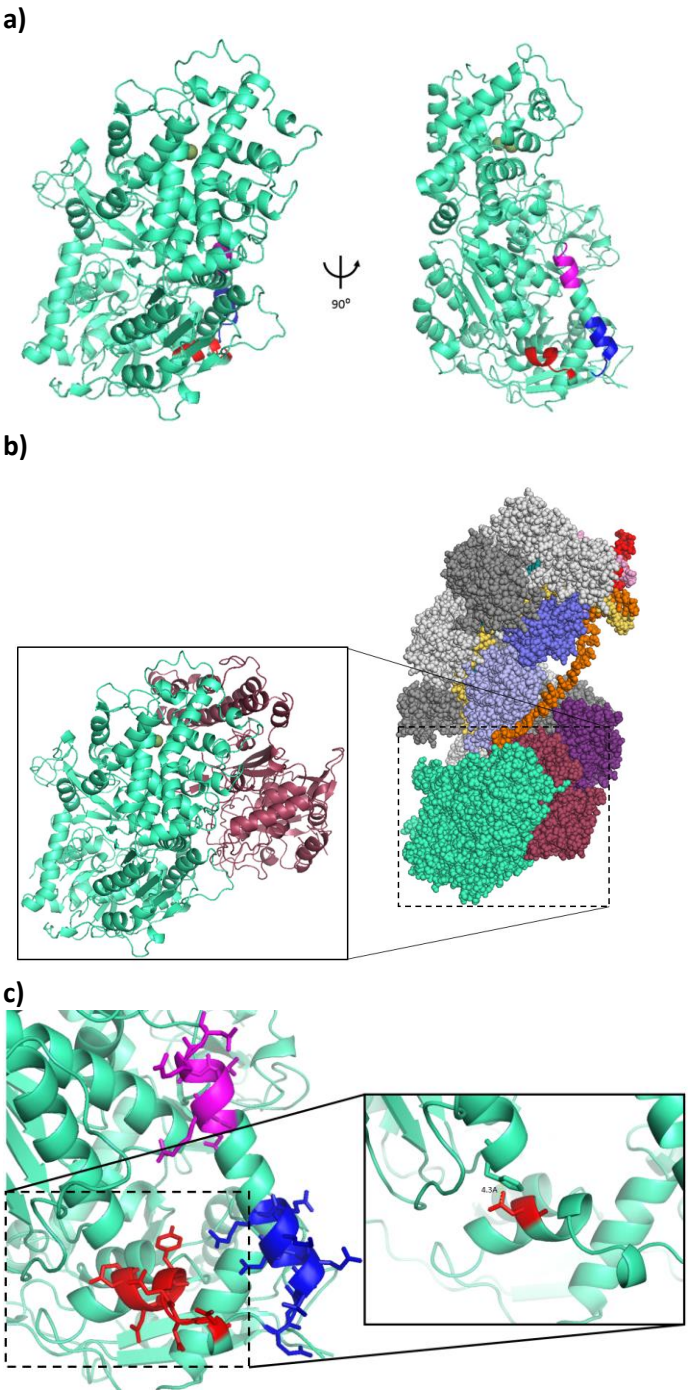


Figure 2.1: Structural overview of Cas3-Cascade interface. **a)** Overall Cas3 structure with linker helix interface residues investigated in this study. AA766-770 (Region 1) in red, AA776-787 (Region 2) in blue, AA791-797 (Region 3) in magenta. **b)** Structural model of cascade-Cas3 association (from Xiao *et al.* 2018) (57). **c)** Zoom-in of the linker helix, showing interacting residue orientations. Interaction between D766 and H487 shown in the inset.

2.3.1 Region Mutations

Two rounds of mutagenesis were carried out to narrow the mutational effects from the entire Linker Helix region to single amino acid residues. The helix deletion was first divided into three smaller deletions, covering a region just outside the helix to the N-terminal side, the N-terminal end of the helix, and the C-terminal section of the helix (Figure 1c). The mutants were assayed for Cas3-binding defects using electrophoretic mobility shift assays (EMSA) and Cas3-mediated DNA cleavage defects analyzed by denaturing PAGE.

In the EMSA experiments, increasing concentrations of Cas3 were added to Cascade bound to a fluorescent substrate in a pre-locked R-loop conformation. Comparisons were made between the mutant cas3 affinity for this Cascade-Substrate complex and the wild-type affinity. A loss of the up-shifted band which results from Cas3 binding is interpreted to be the result of a disruption of the Cas3-Cascade interaction.

In the DNA cleavage experiments, Cas3 mutants were assayed for their ability to degrade substrate DNA. Again, the Cascade was pre-loaded with fluorescently labelled substrate. Cas3 was then added to this complex and heated to a reactive temperature (58C) in the presence of excess Cobalt and ATP. Wild-type Cas3 degrades the substrate in a dose-dependent manner. Since Cas3 is only able to degrade double-stranded DNA when recruited to a fully R-looped Cascade, this is an indirect read-out for Cas3's ability to associate to Cascade in a functionally relevant manner. The processive cleavage products of the mutants were visually assessed for cleavage efficiency compared to wild-type.

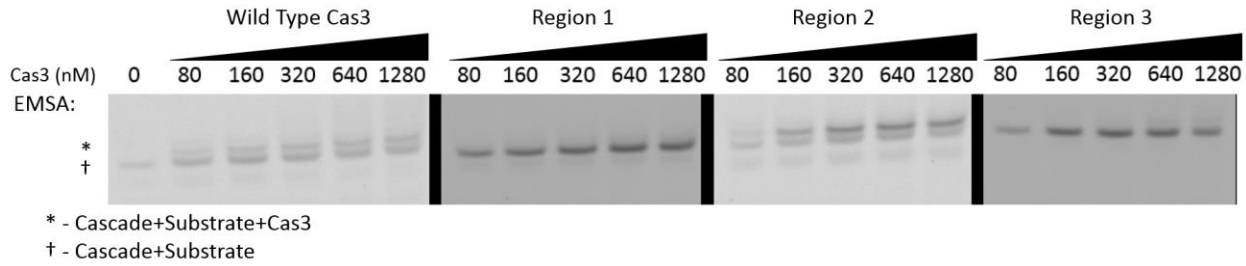
Those mutants which had five or more amino acid substitutions were included in the category of Region Mutations. These regions were AA 766-770 (covered as the mutant called "Region 1" – this region is outside the Linker Helix at the N-terminal end), 776-787 (covered in the mutant "Region 2" – this region is comprised of residues at the N-terminal end of the linker helix), and 791-797 ("Region 3" at the C-

terminal end of the linker helix) **(Table 2.1)**. The Region 1 and Region 2 (766-770 and 776-787) were targeted as glycine mutations because they were outside of the helix, presenting looping structures along the surface of Cas3. In Region 2, an arginine was chosen to replace an aspartic acid at position 781 because of an interaction with R742 of the Helicase domain, with the logic being that this would affect a maximal disruption. Region 3 was a standard alanine scan substitution. Region 1 and Region 3 present an easily interpreted binding and cleavage defect – we believe it is clear that these mutants are disrupting binding and thus severely affect cleavage **(Figure 2.2)**. Region 2, however, displays a phenotype of increased binding in EMSA. Surprisingly, despite this result, the cleavage phenotype for this mutant is as severe as the most severe cleavage defects **(Figure 2.2)**.

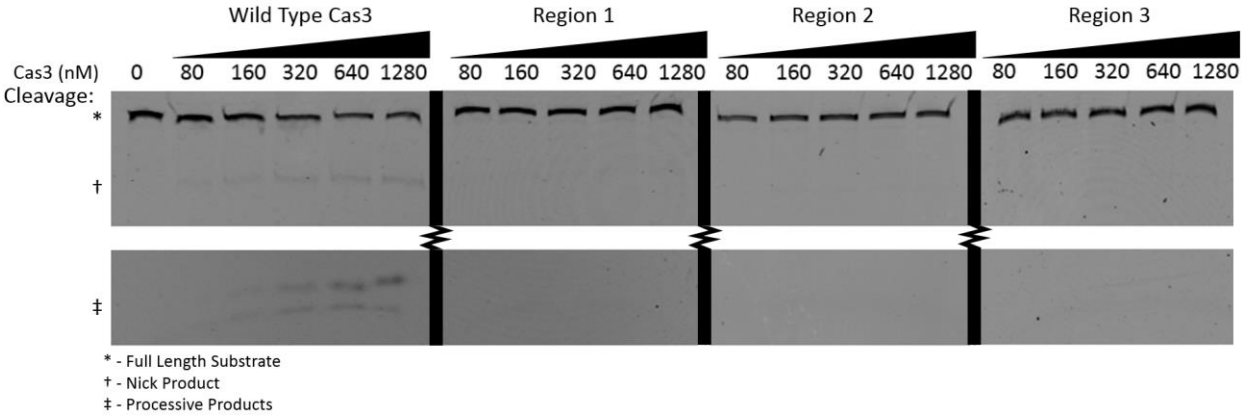
Table 2.1: Summary of Cas3 mutants and their effects on biochemistry. Amino Acids mutated, given names for the paper, effects on binding, and effects on cleavage.

Mutant	AA Mutated	Binding	Cleavage
Wild Type	/	**+**	**+**
Region 1	D766G, D767G, D770G, D771G, D772G	**- - -**	**- - -**
D766A	D766A	**- - -**	**- -**
770-772	D770A, D771A, D772A	**- -**	**- - -**
Region 2	E776G, D777G, D781R, E783G, R784G	**++**	**- - -**
776-779	E776A, D777A, L778A, E779A	**-**	**- -**
781-784	D781A, M782A, E783A, R784A	**-**	**- -**
Region 3	Q791A, R792A, L794A, A795G, R796A, N797A	**- -**	**- - -**

a)



b)



c)

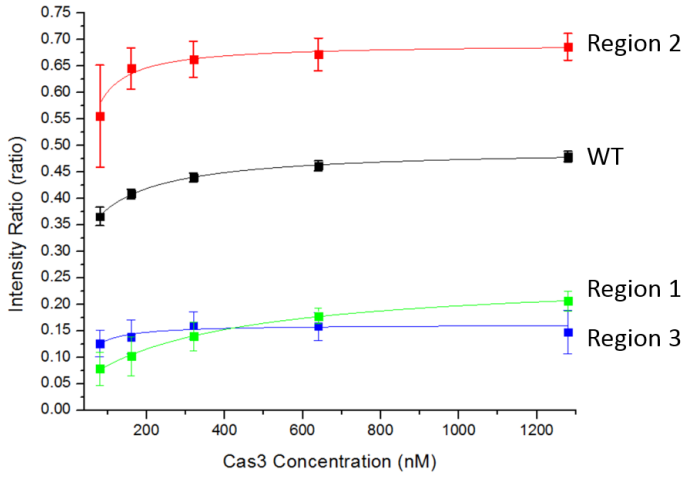


Figure 2.2: Cas3 Region mutants data for Region 1, Region 2, and Region 3. **(a)** Agarose native gel showing Cascade-bound substrate (†) and Cas3-Cascade-bound substrate (*) over Cas3 titration scanned for 6FAM, quantified in (c) **(b)** Urea dPAGE cleavage experiment on the same substrate as (a) showing nick and processive cleavage products in the presence of ATP. All mutants show no significant cleavage products. **(c)** quantification of native gel showing significant binding defects for Region 1 and Region 3 mutants. Surprisingly, Region 2 shows an increase in binding, but abrogation of cleavage (b).

From these results, we can make a few preliminary observations and predictions. The residues at the most N-terminal end of the linker helix are not directly involved in a specific binding interaction since the Region 2 mutant exhibited strong binding to Cascade *in vitro* but no detectable cleavage activity on DNA substrates. This conclusion is supported by the high-resolution Cryo-EM structure (57). The mutant at the C-terminal end of the helix (Region 3) is more readily interpretable in these results since mutation resulted in coupled binding and cleavage defects. We could also start to form a hypothesis that the helicase may be important for proper Cas3 binding in the Region 1 mutant. However, from these experiments alone, it is not possible to determine whether this restriction on binding is due to modification of the binding surface or due to restriction of a necessary conformation in the helicase.

2.3.2 Specific Mutations

To further investigate key residues and the important features of Cas3 which are necessary for association to Cascade, the above region mutants were split into smaller, more targeted sections. The conserved negative charges were specifically targeted based on the hypothesis that the Cascade-Cas3 interaction may be mediated by favorable electrostatic interactions. Region 1, which is outside of the Linker Helix to the N-terminal end, was split into a point mutation at residue 766, substituting alanine for the aspartic acid at that position. Additionally, another mutant covering the stretch of aspartic acid residues along the surface of Cas3 at AA 770-772 was generated as three alanine substitutions. Additionally, these negative residues make a number of favorable interactions with positive residues from each of the helicase domains (H487 and R749). In our assays, 770-772 and D766A elicited severe phenotypes in both binding and cleavage **(Figure 2.3)**. D766A is the only point mutant to show a phenotype, with no detectable binding or cleavage whatsoever.

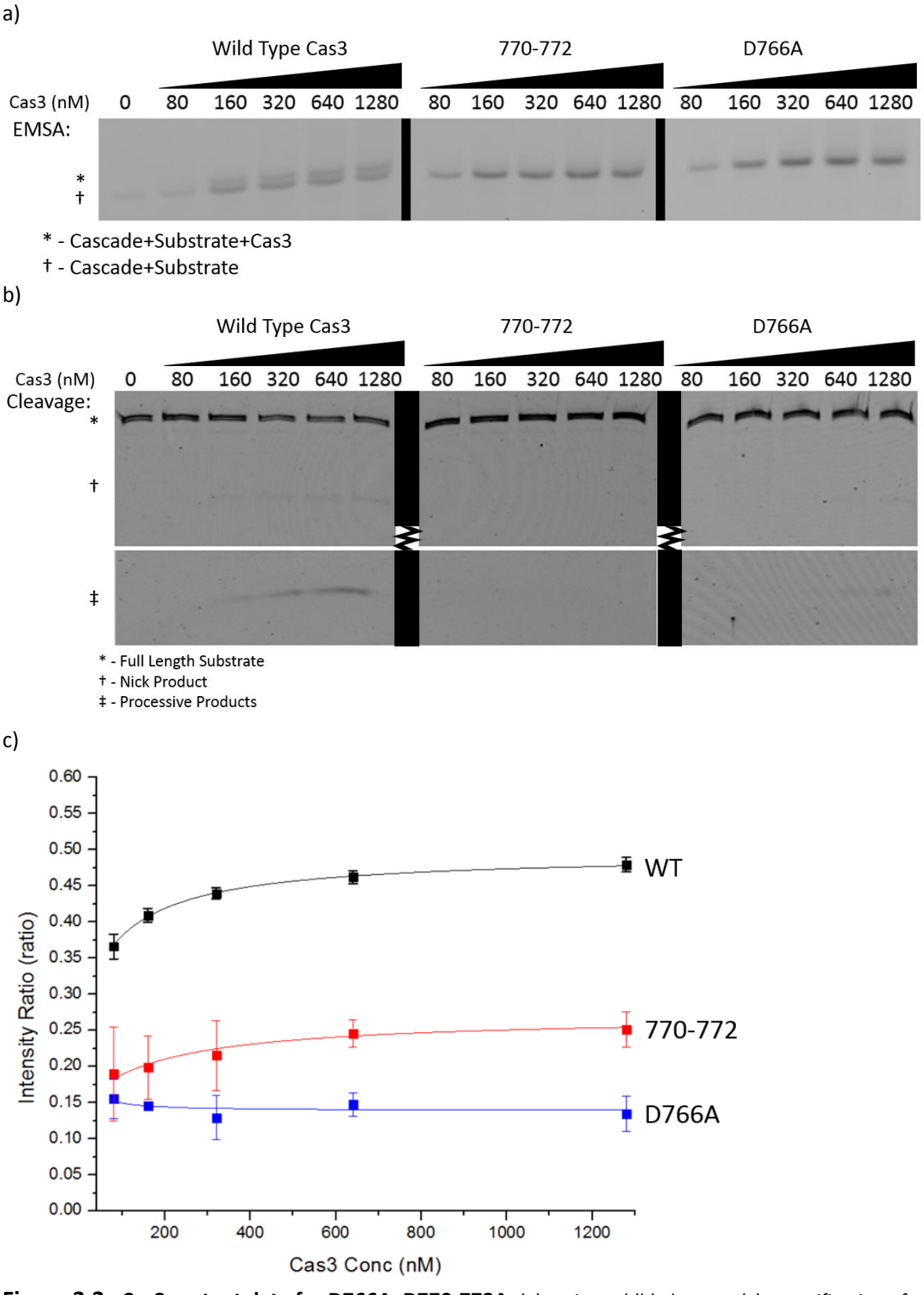
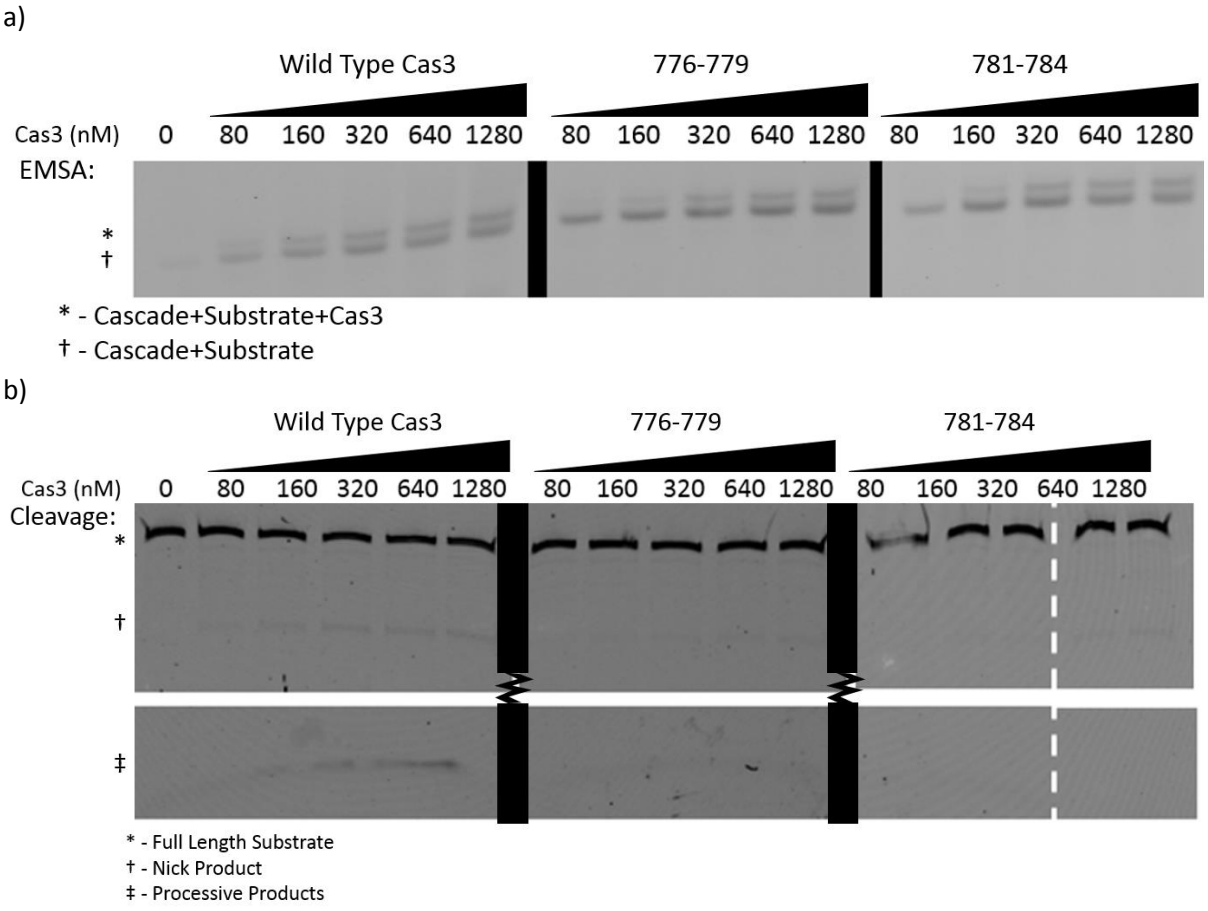


Figure 2.3: Cas3 mutant data for D766A, D770-772A. (a) native gel (b) cleavage (c) quantification of native gel. Both D766A and D770-772A mutants show a decrease in binding affinity and abrogation of cleavage activity.

Another set of targeted mutants included AA positions 776-779 and 781-784 which comprise the N-terminal section of the Linker Helix. These two mutants constitute a split of Region 2 and were performed as an alanine scan. Because of Region 2's surprising characteristics, these mutants are necessary for understanding the properties of this segment of the Linker Helix. Binding and cleavage phenotypes for these two mutants were more intermediate than Region 2 **(Figure 2.4)**. 776-779 and 781-784 still disrupted cleavage considerably, though both had a moderate effect at most on binding.

In an attempt to extend this information and provide more information on the specific residues involved in the interaction surface, single alanine substitutions were performed at various positions on the helix. Point mutations at positions R796, N797, L789, E783, M782, E779, E776, and D777 were generated and tested, none of which showed an appreciable phenotype in a pilot cleavage experiment **(data not shown)**.



c)

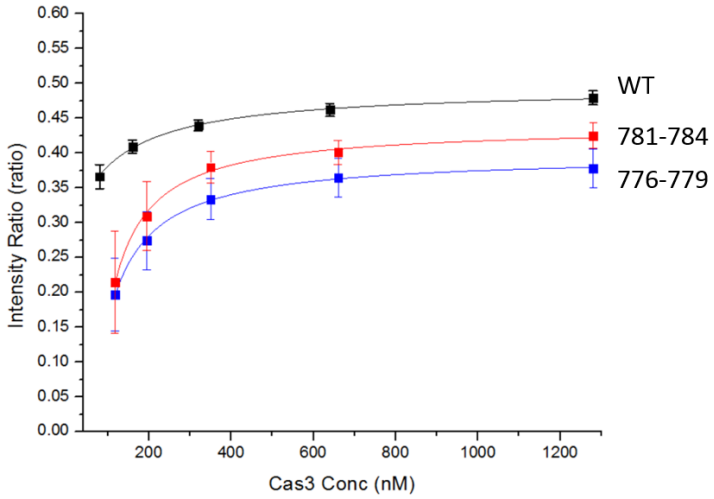


Figure 2.4: Cas3 mutant data for 776-779, 781-784. (a) native gel (b) cleavage (c) quantification of native gel. Both 776-779 and 781-784 mutations showed a moderate decrease in binding affinity but an abrogation of cleavage activity.

2.4 DISCUSSION

In mutating Cas3, the resilience of the interaction to disruption by point mutation along the Linker Helix reflects a complex interaction all along the length of the Linker Helix such that disruption of any individual interaction is not sufficient to cause a detectable binding defect. We have shown some mutants which demonstrate only a moderate Cascade binding defect while displaying a severe cleavage defect (such as mutations from 776-779 and 781-784). This implies that there is a set of interactions which are required to validate the association of Cas3 through contact signaling between Cas3 and Cse1. Alternatively, it may be that the geometry of this interaction is so delicate that disruption of or alteration of the orientation in the interaction surface even slightly results in changes to the placement of the helicase and/or nuclease domains. This complex set of interactions would explain well why solvent-exposed point mutations along the linker helix are not sufficient to induce binding or cleavage defects and may also point toward the specificity of the Cas3-Cascade interaction such that Cas3 can only interact with Cascade from the same organism[4]. Variability in sequence conservation along the linker helix supports this conclusion **(Figure 2.5)**.

[4] Personal communications.

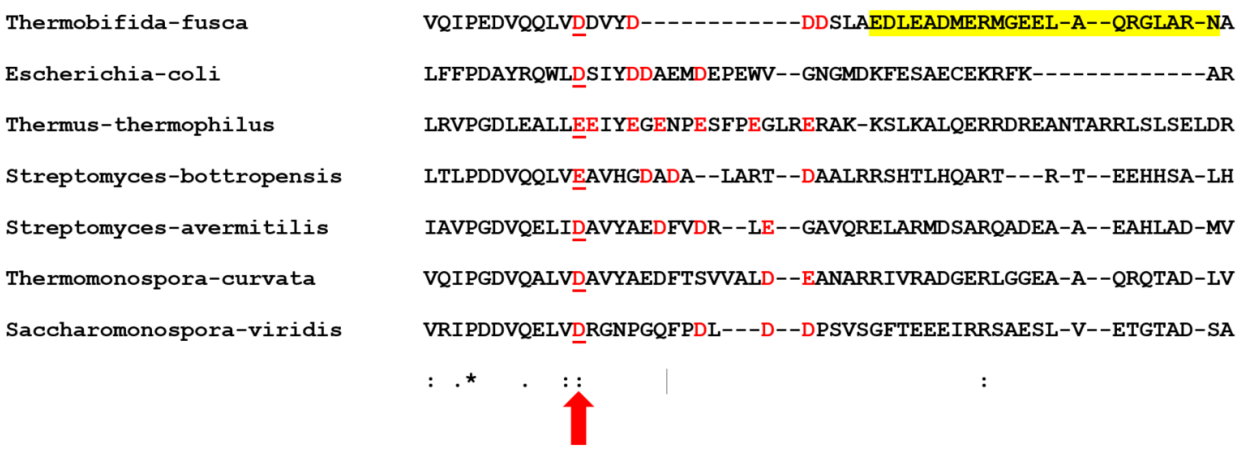


Figure 2.5: Cas3 linker helix sequence alignment. The linker helix and surrounding region to the N-terminus is not very strongly conserved across species which utilize the Cas3-Cascade Type I-E CRISPR system. The linker helix is highlighted in the *T. fusca* sequence. D766 (red arrow) is strongly conserved, pointing towards a conserved interaction to the RecA domains.

In an attempt to use some information gained by the experiments reported here, we performed a series of mutations on Cse1 to attempt to perturb the interaction from the other side. Several targets were selected, but in the end through a pilot cleavage screen, none showed a cleavage defect which was separable from non-specific substrate binding defects. Possible targets which we investigated in *T. fusca* Cse1 were R137, R139, R205, R303, and the L3 loop (Y315-P325). All of these mutations were targeting prominent surface features of Cse1 which had solvent exposed positive charges.

Allosteric Effect of D766A on Cas3 Association

Based on the high resolution Cryo-EM structure, it is probable that the D766A mutant phenotype is best described as an allosteric effect (57). If there were a cleavage defect only and no binding defect, it could effectively be argued that this mutation constitutes a structural mutant which disrupts the activity of the helicase. However, given the direct measurement of binding in the EMSA experiments, it is clear that there is a stark binding defect which stands on its own. The positioning of D766 over the two RecA domains of the Helicase lends further credence to this suspicion that it may be altering the surface structure (and thus the geometry of residues necessary for binding) or limiting accessibility to a necessary conformation for the helicase to bind. D766 forms a favorable interaction along the surface of Cas3 to the backbone amine group of H487, bridging the two lobes of the helicase. Additionally, other aspartic acids in this region form similar interactions with the helicase, such as D770 to H487 and D771 to R749. These interactions may play a role in stabilizing a hinge motion on the helicase in this region as a part of a necessary conformational change for Cas3 recruitment and activity. Based on the high level of sequence conservation at this position across Cas3 proteins from different bacteria (generally either D or E amino acid identities), we predict that this effect on Cas3 association is a conserved general mechanism.

Effects of Region 2 on Cas3 Association

We believe the Region 2 mutant, which affected residues at the N-terminal end of the linker helix, displays important and nuanced information on the mechanism of Cas3-Cascade interaction. Two things are true of this mutant phenotype: Binding still occurs but cleavage is strongly affected. This result is surprising and non-intuitive, but it implies several things. First, we can assume that most of the residues responsible for binding are not largely affected – a conclusion supported by the high resolution Cryo-EM structure (57). We can be relatively safe in our assumption because the severity of the Region 2 mutant is drastic in both its scale and in the chemical property changes to that segment of the linker. We have removed four charged residues from this region of the helix and replaced them with glycine while simultaneously performing a charge swap in the middle of this region (D781R). We believe that Region 2 is not directly involved in binding. This assumption is supported somewhat by the intermediate phenotypes present in the 776-779 and 781-784 mutants in which even the non-solvent exposed residues were mutated as well, but binding was not strongly affected relative to wild-type. These two mutants displayed weaker Cascade binding than Wild Type which is surprising. I predicted that splitting this region up should result in an intermediate phenotype between Wildtype and Region 2. What may instead be going on, is that by mutating the residues in Region 2, we have forced the helix to adopt a binding-competent conformation constitutively or otherwise have made the binding-competent conformation more accessible energetically due to increased flexibility in this region. One explanation for how this conformation change is affected is through the Arginine substitution at position 781 creating an intramolecular contact with the stretch of Aspartic Acid residues in the span of 770-772 which is nearby.

To explain the lack of cleavage in the Region 2 mutant, we propose two possible models. First, that this region is important for either sending or receiving a signal to enact the hand-over of substrate DNA from Cascade to Cas3. In this model, there is a signal sent through a contact in this region to validate the

correct association of Cas3 and initiate the dissociation of the DNA target from Cse1. Afterwards, the dissociated target DNA is loaded onto Cas3 for processive cleavage. The fact that both 776-779 and 781-784 maintain severe cleavage defects while having a smaller mutated range in this region is consistent with this model.

As an alternative explanation, it may be possible that this mutation has introduced a change in the orientation of the helix relative to the rest of Cas3 while leaving residues responsible for binding intact. This model would explain the lack of cleavage through the nuclease and helicase being out of place relative to Cse1 because of the change in orientation on the helix. If this were true, it still supports the conclusion that the linker helix is an important anchor point for a complex Cas3-Cascade interaction.

MATERIAL AND METHODS

Cloning: Cse1, CasB-E, and Cas3 Wild-type coding sequences were amplified from *T. fusca* whole genomic DNA and restriction cloned into pET19b, pCDF and pSUMO respectively with an N-terminal strep-tag. CrRNA was generated as a gene synthesis product and restriction cloned into pRSF. These plasmids were transformed to BL21 competent cells for expression (Cse1, CasB-E, and crRNA triple transformed and Cas3 expressed separately).

Protein Purification: Terrific Broth **(**TB - VWR) cultures with antibiotics (Kanamycin: 100ug/mL, Ampicillin: 100ug/mL, Spectinomycin: 50ug/mL) were inoculated with 5-25mL of overnight LB starter culture and brought up to OD ~0.6 in a 37C shaker. Temperature was changed to 18C once the culture reached desired density, and 1mM IPTG was added to induce expression overnight. Cells were spun down at 4000 RPM, re-suspended in Lysis Buffer (500mM NaCl, 50mM HEPES buffered to pH 7.5 for Cas3 and 150mM NaCl, 50mM HEPES pH 7.5 for Cascade), and lysed via sonicator. Whole cell lysate was spun down at 15,000 RPM and the supernatant removed to a separate flask on ice. Strep resin was regenerated using HABA buffer (1mM 2-[4'-hydroxy-benzeneazo] benzoic acid corrected to pH ~8.5) and

HABA was eluted with 20mM NaOH. NaOH was washed off with Lysis Buffer to equilibrate the column. Supernatant was flowed over the column and washed three times with 3-4 resin volumes of Lysis Buffer. 15-20 mL of Elution Buffer (2.5mM desthiobiotin in Lysis Buffer) was flowed over the washed resin beads and the flow-through collected. Protein in Elution Buffer was concentrated in a 10,000 MW spin column concentrator to <1 mL total volume and diluted with clean Lysis Buffer without desthiobiotin, then concentrated again in the same spin column to a similar volume to remove residual desthiobiotin. Concentrated protein was then loaded to an AKTA-pure SD200 SEC Column (GE Healthcare) in Sizing Buffer (20mM Tris, 150mM NaCl, 2mM DTT). Fractions were collected at a rate of 1mL/min and the peak of the peak was selected for assay use (Cas3) or the immediate right shoulder of the peak (Cascade). Protein was concentrated in a clean 10,000 MW spin column concentrator to >10 uM concentration then diluted to 10 uM for Cas3 or concentrated to >5uM and diluted to 5 uM for Cascade in Sizing Buffer. 20-40 uL aliquots were prepared and flash frozen using liquid nitrogen and stored at -70C for use in assays. For indicated native gels, Cascade was not frozen but instead diluted to 5uM and stored at 4C for up to a month.

Mutagenesis: Mutant Cas3 and Cse1 proteins were produced via mutagenesis PCR amplified with lab-purified iProof DNA Polymerase. Each PCR reaction followed the suggested protocol from NEB (Protocol M0530). PCR reactions were digested with 2uL of DpnI restriction enzyme from NEB and incubated at 37C for ~2hrs. DpnI digested products were then spin-column purified using a Thermo Scientific GENEJet PCR Purification Kit. Products were eluted in ~30 uL of MilliQ water. Ligation reactions were set up with 11.5 uL of spin-column purified products, 1.5 uL 10x T4 Ligase buffer, 1 uL Polynucleotide Kinase, and 1 uL T4 Ligase (all enzymes and buffers from NEB) and incubated for ~1 hr on the bench. 3 uL of ligation reaction was transformed to lab-produced DH5-Alpha chemically competent cells or BL21 chemically competent cells and plated on the appropriate antibiotic-resistant LB agar plates. Single colonies were selected and inoculated into a 5 mL LB culture, grown at 37C for ~8-12 hrs. Plasmid

purification was performed using a Thermo Scientific GENEJet Plasmid Mini-Prep Kit. Sequencing was performed at the Cornell Sequencing Facility using a universal T7 reverse primer for Cas3 mutants or custom primers for Cse1 mutants (included in supplemental). The plasmid was re-transformed to new BL21 competent cells.

Cas3 Native Binding and Cleavage Assays: A Cascade Substrate Loading reaction was prepared (7.5 nM Fluorescent-labeled substrate (Table 2.2), 40nM Cascade, 50mM HEPES pH 7.5, 150mM NaCl, 5% Glycerol) and incubated for 40 minutes at 58C in a thermocycler. Cas3 was prepared in a 2560 nM stock and diluted to 160, 320, 640, 1280, and 2560 nM. An equal volume of Cascade Loading Reaction product was mixed with the Cas3 solutions and incubated overnight at 4 C. A 2% Agarose gel was prepared with TA Buffer (40mM Tris, 20mM Acetic Acid, pH corrected to ~8.2). Running buffer was pre-cooled and the gel was run at constant power (11W) and the gels were loaded at 4 C while current was flowing. Total run-time for the gels under 11W was 3hr. Gels were then scanned using a Typhoon Scanner for Cy5 or FAM tags and the better resolution between the two was used for quantification.

Table 2.2: Fluorescent Substrate used for cleavage and EMSA experiments AAG indicates the PAM, underlined sequence is the Cascade target site. The average Cas3 nicking site is between the two nucleotides indicated in RED.

Description	Sequence
Substrate Full Sequence	taatacgactcactatagggGAATTGTGAGCGGATAACAATTCCCCTGTAGAAAT AATTTTGTTTAACTTTAATAAGGAGATATACCATGGGCAGC**AGCCATCACCATCA TCACCA**CAGCCA**GGATCC**AAGCCAGTGATAAGTGGAATGCCATGTGGGCTGTC**CT CGAG**TCTGGTAAAGAAA**CCGCTGCTGCGAAATTTGAA**CGCCAGCACATGGACTCG TCTACTAGCGCAGCTTAATTAACCTAGGCTGCTGCCA**CCGCTGAGCAATAACTAG C**
Substrate Forward Primer	5'FAM-taatacgactcactatagggg
Substrate Reverse Primer	5'Cy5-gctagttattgctcagcgg

For cleavage/native assays 320, 640, 1280, 2560, and 5120 nM Cas3 aliquots were prepared and diluted in 2x Binding Buffer (final concentrations: 50 mM HEPES pH 7.5, 150 mM NaCl, 5% Glycerol). 20 uL of each of the Cas3 solutions was combined with 40uL of the Cascade training solution. From these solutions, two aliquots were prepared: 30uL for the cleavage reaction, 15uL for the native binding reaction. The cleavage reaction received a cleavage supplement to bring the cleavage reaction to a final composition of 1x Binding Buffer, 20nM Cascade, 3.75 nM substrate, 2 mM ATP, 10 uM $CoCl_2$, 10 mM $MgCl_2$ and Cas3 concentrations from 80 to 1280 nM. The 15 uL aliquots for the Native Binding assay were then corrected with 5 uL of 1x Binding Buffer to a final concentration of 20 nM Cascade and Cas3 concentrations from 80 to 1280 nM. These native binding reactions were then treated identically to the Native binding assays described above. Cleavage reactions were incubated in a water bath at 58C for 40 minutes and the samples were frozen overnight. The next day, an equal volume of Phenol solution was added and the reactions were vortexed. The aqueous phase was taken and an equal volume of 95% formamide solution was added, then the samples were heated to 98C for 5 minutes in a thermocycler. The samples were run on a pre-warmed 8% Acrylamide (19:1) 0.5x TBE urea gel for 40 minutes under 25W of constant power. Afterwards, the gels were allowed to cool and then were scanned for Cy5 and FAM fluorescent tags using a Typhoon Scanner.

AUTHOR CONTRIBUTIONS

A.D., Y.X., and A.K. designed the experiments and interpreted the data. A.D. performed the experiments and wrote the manuscript.

CHAPTER III CREDITS

Credits:

This chapter was adapted from the current draft of our manuscript of working title "Introducing a spectrum of long-range deletions in human embryonic stem cells using Type I CRISPR-Cas", which has been submitted to Cell. As mentioned in the "Author contributions" section: Protein optimization and purification was performed by me. The hESC editing experiments were performed by Zhonggang Hou in Yan Zhang's lab in the Department of Biological Chemistry at University of Michigan. Lesion analysis was performed by Max Gramelspacher and Zhonggang Hou. Sara Howden developed the hESC reporter line used in the experiments. Peter Freddolino performed informatics analyses of the lesions. The experimental design was developed by Ailong Ke, Zhonggang Hou, Yan Zhang, and myself.

CHAPTER III

Introducing a spectrum of long-range deletions in human embryonic stem cells using Type I CRISPR-Cas

Authors: Adam Dolan[1,*], Zhonggang Hou[2,*], Max J. Gramelspacher[2], Sara E. Howden[3,4], Peter L Freddolino[2,5], Ailong Ke[1,#], Yan Zhang[2,#]

Affiliations:

[1] Department of Molecular Biology and Genetics, Cornell University, 253 Biotechnology Building, Ithaca, NY 14853, USA.

[2] Department of Biological Chemistry, University of Michigan, 1150 W. Medical Center Drive, Ann Arbor, MI 48109, USA.

[3] Murdoch Children's Research Institute, Flemington Rd, Parkville, VIC 3052, Australia

[4] Department of Paediatrics, University of Melbourne, Parkville, VIC 3052, Australia

[5] Department of Computational Medicine and Bioinformatics, University of Michigan, 100 Washtenaw Avenue, Ann Arbor, MI 48109, USA

*These authors contributed equally to the work.

[#]Correspondence: ailong.ke@cornell.edu, yzhangbc@med.umich.edu

3.1 Abstract

Powerful genome editing tools derived from single-component CRISPR-Cas systems have revolutionized biomedical research. More sophisticated CRISPR systems exist, with novel enzymatic properties. Here we demonstrate the feasibility of using Type I CRISPR-Cas to effectively introduce a spectrum of long-range deletions in an RNA-guided fashion in human embryonic stem cells (hESCs). Type I CRISPR-Cas relies on the multi-subunit ribonucleoprotein (RNP) complex Cascade to identify a DNA target, and the helicase-nuclease enzyme Cas3 to degrade DNA processively. When delivered as RNP, Cascade and Cas3 introduced a variety of long-range deletions in human genome, ranging from a few hundred nucleotides to more than 35 kilobase-pairs, in regions upstream of the RNA-guided target site. These results highlight the potential utility of Type I CRISPR-Cas for the functional dissection of coding and non-coding regions in eukaryotic genome, and possibly in erasing large deleterious genetic elements and introducing long-range epigenetic modifications.

3.2 Introduction

The majority of prokaryotes rely on CRISPR-Cas systems to establish adaptive immunity against foreign genetic elements (33, 173-177). The diverse set of CRISPR-Cas systems rely on the basic principle of an RNA guide complexed with one or more proteins. CRISPR-Cas systems are divided into two major classes: Class 1 systems utilize a multi-subunit effector complex to search and destroy nucleic acid targets, whereas Class 2 systems use a single-component effector complex (178, 179). Each Class is further classified into at least three Types, based on the *cas* operon composition. The utilization of Cas9 (Class 2 Type II) for RNA-guided eukaryotic genome editing revolutionized biomedical research and precision medicine (180-182). Guided by an engineered sgRNA, Cas9 introduces a DNA double-strand break (DSB) at the targeted site, which is typically repaired via the error-prone non-homologous end joining (NHEJ) pathway, leading to nucleotide insertion/deletions (indels) and gene disruption. Homology-directed repair (HDR) can also occur when a DNA template is present, leading to template-guided gene conversion. A wide variety of Cas9-based tools have been invented for high-throughput genetic screening, epigenome modification, and programmable base editing (183). So far, all CRISPR-based eukaryotic gene editing tools (Cas9, Cas12, Cas13) were harnessed from Class 2 systems. The more prevalent Class 1 CRISPR systems remain as untapped resources (178).

The most widespread and diverse form of CRISPR-Cas, the Class 1 Type I system, uses a very different interference mechanism from that of Cas9. Rather than introducing a single DSB at the targeted site, type I systems shred the DNA target processively through a multi-step process. First, a multi-subunit RNP called Cascade (CRISPR associated complex for antiviral defense) uses a CRISPR RNA (crRNA) to recognize a complimentary target flanked by a 5' protospacer adjacent motif (PAM). This results in stable R-loop formation and triggers a large conformational change in Cascade. The helicase-nuclease fusion enzyme Cas3 is then specifically recruited to the R-loop-forming Cascade, nicks the non-target strand (NTS) DNA, and processively degrades its upstream region (PAM-proximal side). Cas3 further

degrades the target strand (TS) DNA, although the detailed mechanism remains unclear. Among different subtypes (I-A to I-G), the best-understood are the Type I-E systems from *E. coli* (184-193) and *Thermobifida fusca* (*Tfu*) (55, 57, 83, 194, 195).

3.3 Results

In this study, we explored the feasibility of achieving RNA-guided genome editing in hESCs using Type I CRISPR-Cas. We chose the *T. fusca* type I-E system for its clearly defined mechanisms relative to other Type I systems and the highly active Cas3 nuclease **(Figure 3.1A)**. However, the optimal growth temperature for *T. fusca* is 55 ˚C and R-loop formation by *Tfu*Cascade exhibits a strong temperature dependence (195), which presents a potential technical hurdle for its adaptation to mammalian systems. Although robust interference activity was previously observed from *T. fusca* Type I-E CRISPR at 37 ˚C *in vivo* (194), as a precaution, we screened a number of structure-guided mutations aimed at weakening the thermo-adaptation features in *Tfu*Cascade. We found that *Tfu*Cascade bearing a N23A mutation in the Cse2 subunit (195) was more specific in DNA-binding and equally efficient in R-loop formation at mesothermic temperature **(Fig. 3.2A)**. More importantly, this mutant was more efficient in recruiting *Tfu*Cas3 for DNA nicking and degradation at 37 ˚C **(Fig. 3.2B)**. To assay for genome editing activity (**Fig. 3.1B**), we created a hESC dual reporter line (H9-DNMT3B-tdTomato/EGFP) bearing knock-ins of a tandem dimer tomato fluorescent protein (tdTomato) gene and a enhanced green fluorescent protein (EGFP) gene at the two alleles of the highly expressed DNMT3B locus **(Fig. 3.1C)**. RNA-guided disruption of the EGFP allele would lead to the accumulation of cells expressing tdTomato only, and *vice versa* for tdTomato disruption. Electroporation of *Tfu*Cas3 and *Tfu*Cascade RNP was chosen as the delivery strategy to avoid the potential complications of expressing a multi-subunit RNP complex in hESCs. Nuclear localization signals (NLSs) were appended to the C-terminus of *Tfu*Cas3 and the C-terminus of each of the six Cas7 subunits in *Tfu*Cascade **(Fig. 3.2C-D)** to promote nuclear uptake.

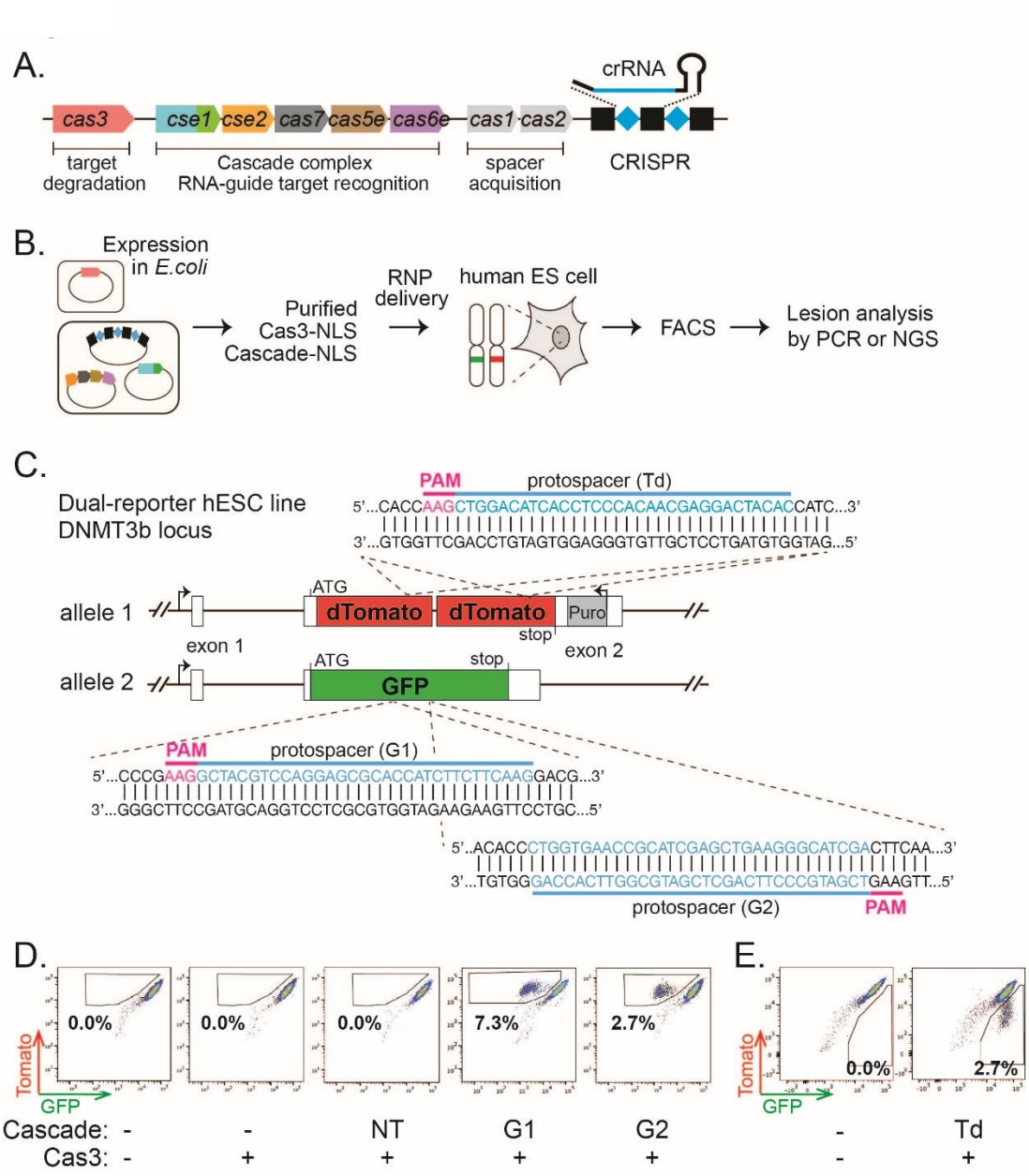


Figure 3.1: Type I CRISPR-Cas can enable RNA-guided genome editing in human ES cells. (**A**) Schematic diagram of the *T. fusca* Type I-E CRISPR-*cas* locus. Black rectangles and blue diamonds represent CRISPR repeats and spacers; colored boxes, *cas* genes. (**B**) Procedure of the genome editing experiments in hESCs. (**C**) Schematic of the hESC dual-reporter line bearing EGFP and tdTomato at the DNMT3B locus. Protospacers for the three reporter-targeting crRNAs are indicated in blue, and corresponding PAMs in magenta. (**D-E**) Flow cytometry analysis of the dual-reporter hESC line 4-5 days after RNP delivery. Percentages of EGFP-negative tdTomato-positive cells are indicated in (**D**), and percentages of EGPF-positive tdTomato-negative cells are indicated in (**E**). Colors in D, E indicate density of cells detected with a certain GFP/tdTomato signal. Cascade NT, G1, G2 and Td indicate Cascade with guides that do not target the genome and are thus non-targeting (NT), target GFP at Guide 1 (G1), GFP at Guide 2 (G2), or tdTomato (Td).

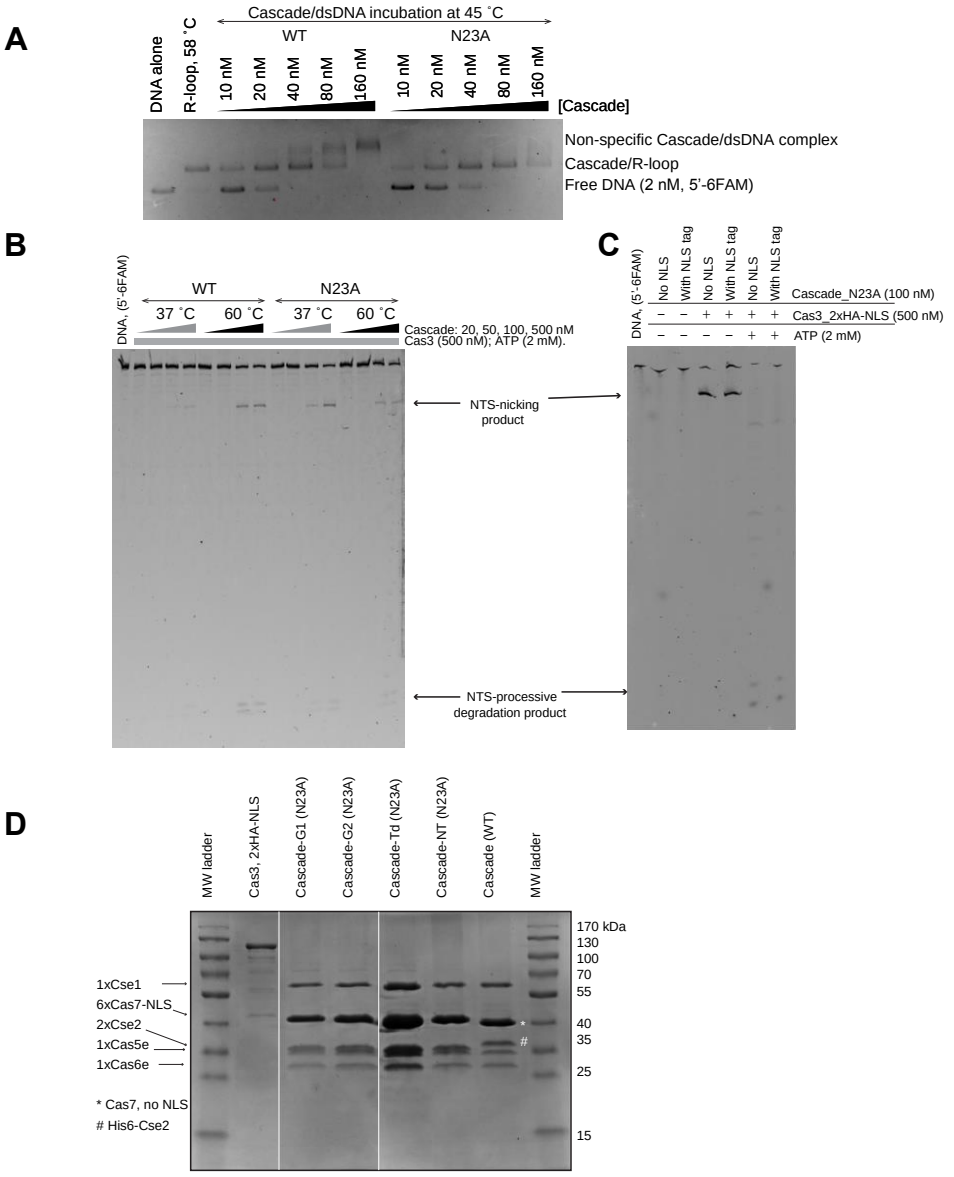


Figure 3.2: Biochemistry on the *T. fusca* Type I-E CRISPR system. (A) R-loop formation behavior of WT and N23A mutant *Tfu*Cascade at 45 ˚C. N23A interacts with dsDNA less non-specifically. **(B)** In comparison with WT, N23A mutant *Tfu*Cascade enables more efficient DNA nicking and degradation by *Tfu*Cas3 at 37 ˚C. **(C)** NLS-tagged *Tfu*Cascade behaves similarly as the untagged version. **(D)** SDS-PAGE analysis of *Tfu*Cas3 and *Tfu*Cascade used in the genome editing experiments. *Tfu*Cascade was programmed with different guide RNAs, as referred in the text.

It has been observed in previous studies (and in Chapter II of this thesis) that the *T. fusca* Cas3 protein has not been very active and significant activity was only observed when supplemented with divalent ions, most notably cobalt, during DNA cleavage experiments (171). This lack of activity in the absence of a metal supplement has also been noted in another thermophilic Cas3 from *T. thermophilus* (196). And this trend extends to *E. coli* Cas3 which also needs to be supplemented with either cobalt or nickel (79, 158). Even so, the *in-vitro* activity when supplemented with divalent ions in the reaction buffer was not very robust. The probable mediator of this decrease in activity is that iron is the native metal ligand for the reaction center of the nuclease, but when recombinantly expressed in *E. coli*, differences between *E. coli* and *T. fusca*'s cellular environment resulted in oxidation of the nuclease iron.

Therefore, it became apparent very quickly that if TfuCas3 could be optimized out of this requirement, it would result in a much higher likelihood of success in gene editing experiments. Towards this goal, I developed a protocol which improves the activity of Cas3 to a very high degree (~35-fold higher activity compared to previous) **(Figure 3.3a)**. The protocol involves starting the expression culture in a small volume of LB overnight (5mL), transferring to a larger starter culture of M9 (100mL) overnight, and then transferring to the expression volume (2L-8L) (see materials and methods). Once the expression culture reaches OD 0.6-0.8, I lowered the temperature to 20°C and added Cobalt Chloride to a final concentration of 100uM. During the expression, using M9 minimal media without a trace metal supplement carefully controls the presence of iron which protects the active site from being poisoned.

Notably, in the presence of ATP, the nicking band in reactions in which the cobalt-containing Cas3 was used is less prominent on the non-target strand **(Figure 3.3b, 3.3c).** This is an indication that processive activity is higher and/or that nuclease activity is stronger, since the fluorescent signal is being spread out over a more diverse set of non-specific cleavage products. Increase in cleavage of the target strand also corresponded with the increase in non-target strand activity.

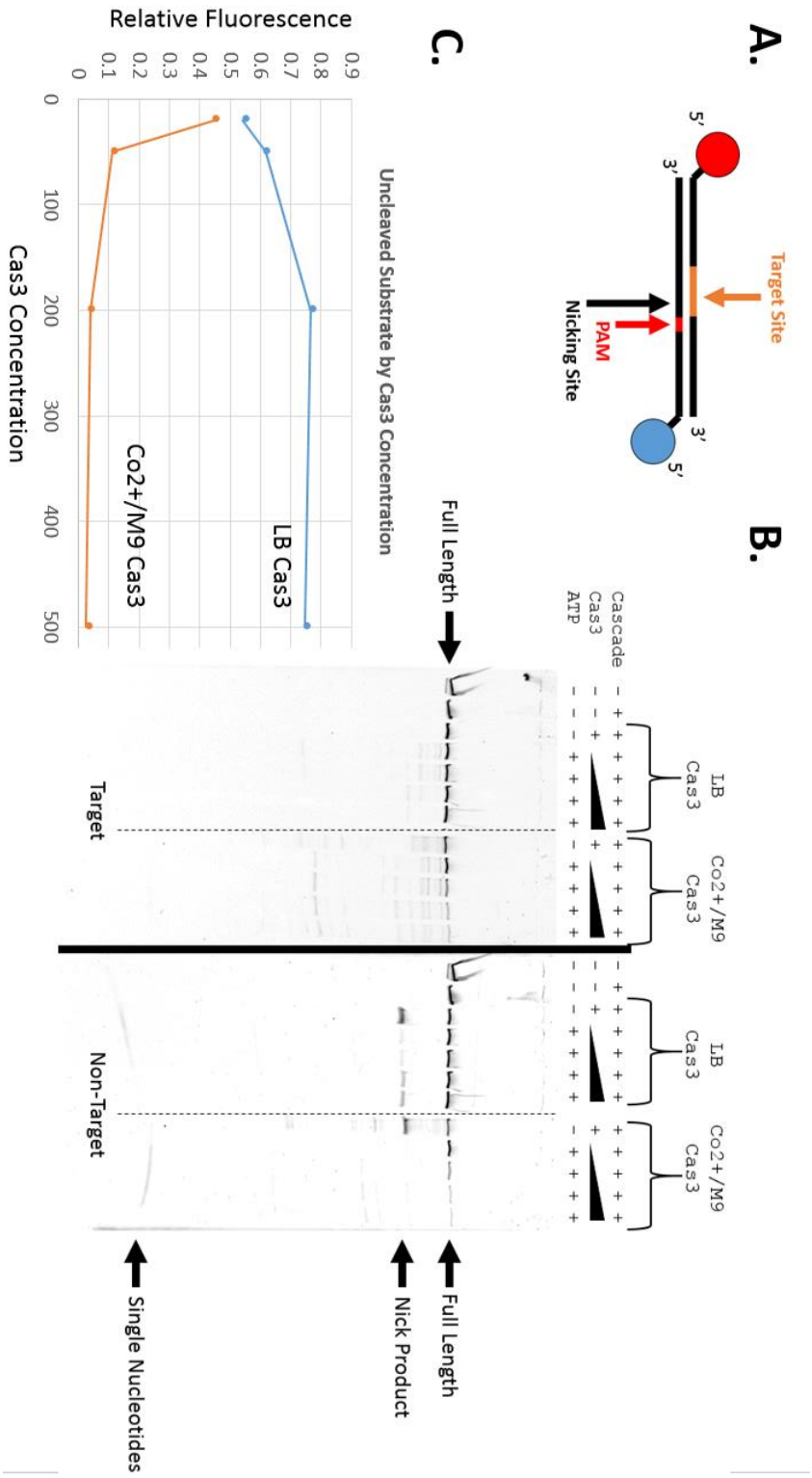


Figure 3.3: Improvement of Cas3 nuclease activity a) Design of the substrate used in the cleavage experiment of panel c. The 5' end of the target strand is labeled with Cy5 (red), the 5' end of the non-target strand is labeled with 6FAM (blue). **b)** Cleavage assay showing increased nuclease activity of Cobalt-purified Cas3. **c)** Quantification of uncleaved substrate in (b) in the presence of ATP expressed as a fraction of fluorescent signal from uncleaved substrate

We programmed *Tfu*Cascade with a crRNA guide (G1) against a 32-bp region in EGFP that was flanked by an interference-enabling PAM (5'-AAG) **(Fig. 3.1C)**, and electroporated it together with *Tfu*Cas3 into the hESC dual-reporter line. A sub-population (7.3%) of EGFP-negative/tdTomato-positive cells were detectable by flow cytometry after 4-5 days **(Fig. 3.1D)**. Negligible levels of EGFP-negative/Tomato-positive cells were detected in control transfections that included a non-targeting (NT) *Tfu*Cascade and *Tfu*Cas3, or *Tfu*Cas3 alone **(Fig. 3.1D)**. To further demonstrate that this novel editing platform is programmable, we evaluated two additional *Tfu*Cascade RNPs. Co-delivery of a *Tfu*Cascade-G2 targeting the opposite strand of GFP (**Fig. 3.1C**) together with *Tfu*Cas3 lead to the accumulation of 2.7 % EGFP-negative/tdTomato-positive cells (**Fig. 3.1D**). Furthermore, electroporation of a tdTomato-targeting (Td) *Tfu*Cascade in conjunction with *Tfu*Cas3 resulted in a 2.7% Tomato-negative/EGFP-positive cell population **(Fig. 3.1E)**. No apparent cell toxicity was noticed for any combination of Cas3 and/or Cascade delivery. A small fraction of cells lacking both EGFP and Tomato fluorescence were observed for each sample, even when no CRISPR components were delivered. This was most likely caused by spontaneous hESC differentiation that led to repression of DNMT3B-driven reporter gene expression. Collectively, these experiments suggest that type I-E CRISPR systems can induce programmable gene disruption in hESCs, and that both the nuclease-helicase Cas3 and a cognate Cascade are required.

The efficiency of EGFP disruption positively correlated with *Tfu*Cascade concentration, increasing from 3.3% to ~13.1% when the amount of *Tfu*Cascade-G1 RNP was increased from 20 to 80 pmole, with *Tfu*Cas3 concentration kept constant at 20 pmole **(Fig. 3.4A)**. Similarly, increasing the amount of transfected *Tfu*Cascade-G2 RNP also lead to enhanced editing efficiency (**Fig. 3.4B**). In contrast, doubling, tripling or quadrupling the amount of *Tfu*Cas3 while keeping Cascade concentration constant did not improve editing efficiency **(Fig. 3.4C)**. These findings suggest that the editing efficiency in hESCs is currently limited by the target-searching efficiency of *Tfu*Cascade rather than the DNA cleavage activity of *Tfu*Cas3.

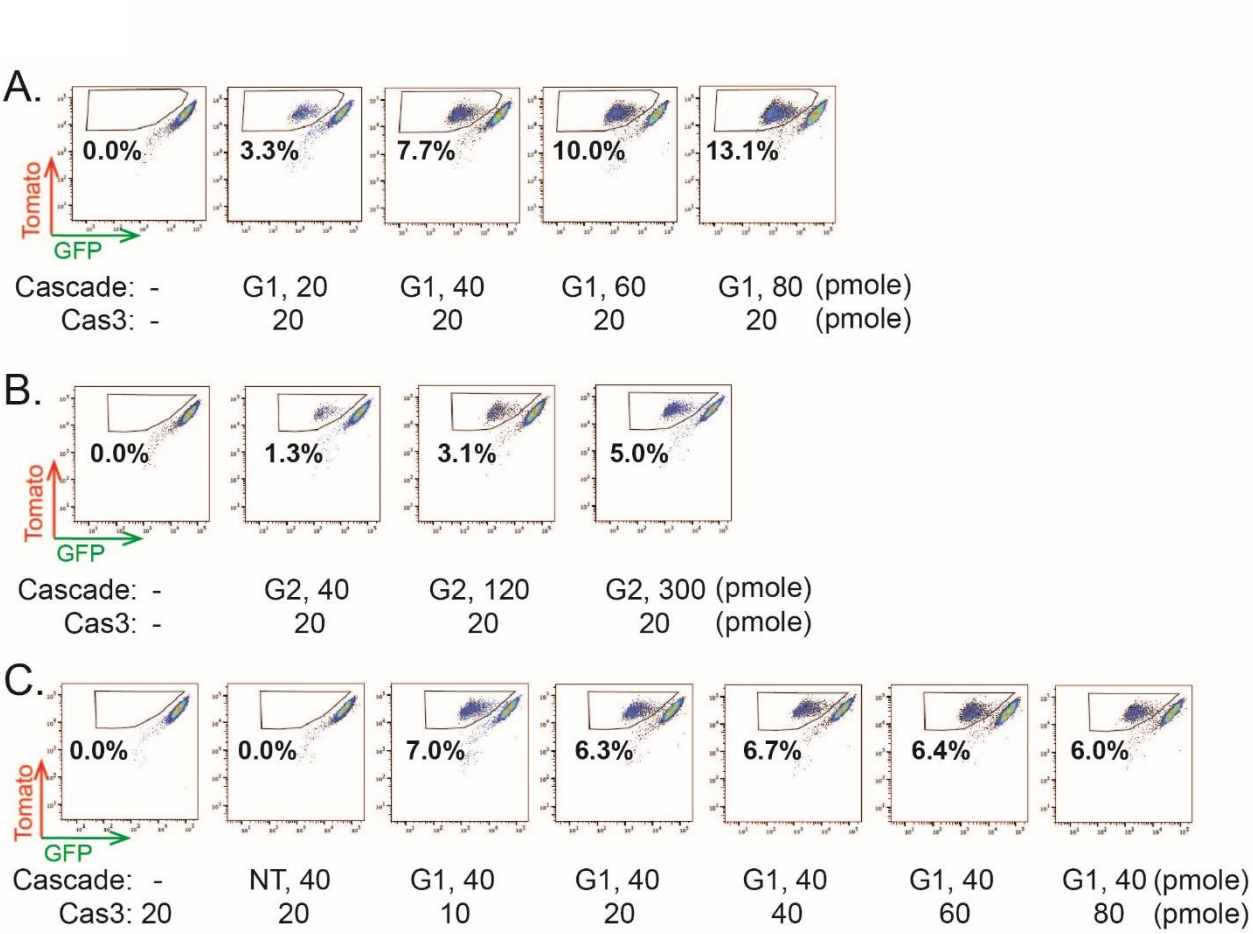


Figure 3.4: Optimization of genome editing efficiency. (**A-C**) Flow cytometry analysis of dual-reporter hESC line 4 days after RNP delivery. Increasing amount of TfuCascade-G1 or TfuCascade-G2 was used in conjunction with constant TfuCas3 in (**A**) and (**B**). Increasing amount of TfuCas3 was used in conjunction with constant amount of TfuCascade-G1 in **(C)**. Percentages of EGFP-negative tdTomato-positive cells are indicated.

Based on prior knowledge of how type I CRISPR cleaves DNA, we speculated that chromosomal deletions were induced upstream (i.e. PAM-proximal side) of the target site. To define these lesions, we extracted genomic DNA from *Tfu*Cascade-G1/Cas3 edited hESCs before and after fluorescence activated cell sorting (FACS), and PCR-amplified a ~5.1 kb region using two primers spanning a region 4.7 kb upstream and 400 bp downstream of the target site **(Fig. 3.5A).** The unedited cells and the *Tfu*Cascade-NT/Cas3 treated cells served as two negative controls; both produced a single PCR band of 5.1 kb, suggesting that the DNMT3B-EGFP locus was intact **(Fig. 3.5B, lanes 1-2)**. The amplicon from the unsorted total cell population after the *Tfu*Cascade-G1/Cas3 treatment contained a faint ladder of smaller bands in addition to the full-length band, indicating that a fraction of these cells harbored deletions of varying lengths at the DNMT3B-EGFP locus **(Fig. 3.5B, lanes 3)**. Notably, PCR amplification from the FACS-sorted EGFP-negative/tdTomato-positive cells were enriched with a distribution of smaller products, ranging from ~5 kb to 1 kb in size (**Fig. 3.5B**, **lane 4**). Speculating that some deletions may extend beyond the 4.7 kb detection limit, we repeated the same experiment using a different forward primer annealing further upstream (~8.5 kb) of the target site (**Fig. 3.5A**). The result revealed that the chromosomal deletions were indeed well-represented all the way to ~7.5 kb (**Fig. 3.5B, lanes 5-8**). Control PCRs amplifying a 5.5 kb region downstream of EGFP revealed no detectable genomic deletions (**Fig. 3.5B, lanes 9-12**). This profile is in stark contrast to the eukaryotic gene editing profiles by the RNA-guided Cas9 or Cas12 nucleases, which typically lead to small indel formation at the target site.

We further attempted to map the precise boundaries of Cascade-G1/Cas3-induced lesions. In a low-throughput approach, we topo-cloned the amplicons from FACS-sorted samples in lanes 4 and 8 of Fig. 3B, and randomly picked fifteen clones for Sanger sequencing to identify the chromosomal junctions. Each clone revealed a unique chromosomal deletion profile upstream of the target site (**Fig. 3.5C, Table 3.1**). These data suggest that the Type I-E CRISPR induces at least two DSBs in this region, and possibly

more in between. In all cases, the 5' and 3' regions flanking the deletions were directly re-ligated, presumably by the host NHEJ pathway. The 5' lesion boundaries, which likely reflect the last DSB generated by Cas3 before dissociating from the DNA, are distributed across a region several kb in size, highlighting the heterogeneous nature of Cas3-induced deletions. An unexpected finding was that the 3' boundaries of these lesions, which represent the first DSB by Cas3, did not line up precisely at the Cascade recognition site. Instead, they spread out stochastically within the 273 nt EGFP coding region upstream of the target site (**Fig. 3.5C**). More editing events likely started further upstream in a long intron preceding the EGFP coding sequence, which may not be detectable in cells enriched by FACS given that EGFP expression may be unaffected. This observation is difficult to rationalize based on the existing mechanistic model, as it suggests that Cas3 is not capable of generating a DSB during the very initial phase of DNA translocation. Previous single molecule biochemistry revealed that after recruitment by Cascade, Cas3 initially reels dsDNA towards itself repeatedly while still associated with Cascade, then eventually dissociates from Cascade and translocates on DNA for kilobases in distance (83); in both phases NTS DNA was sporadically erased, leading to the exposure of TS ssDNA (83, 197). However, DSB formation was not observed at the single molecule level, even though DNA were found in bulk experiments to be shredded into pieces, (83, 197). This discrepancy could not be rationalized in the past. Here our human cell genome editing experiments revealed a surprising DSB formation pattern, suggesting that the current understanding of the DNA degradation mechanism by Type I CRISPR machinery is still incomplete and can be further illuminated by human cell gene editing studies.

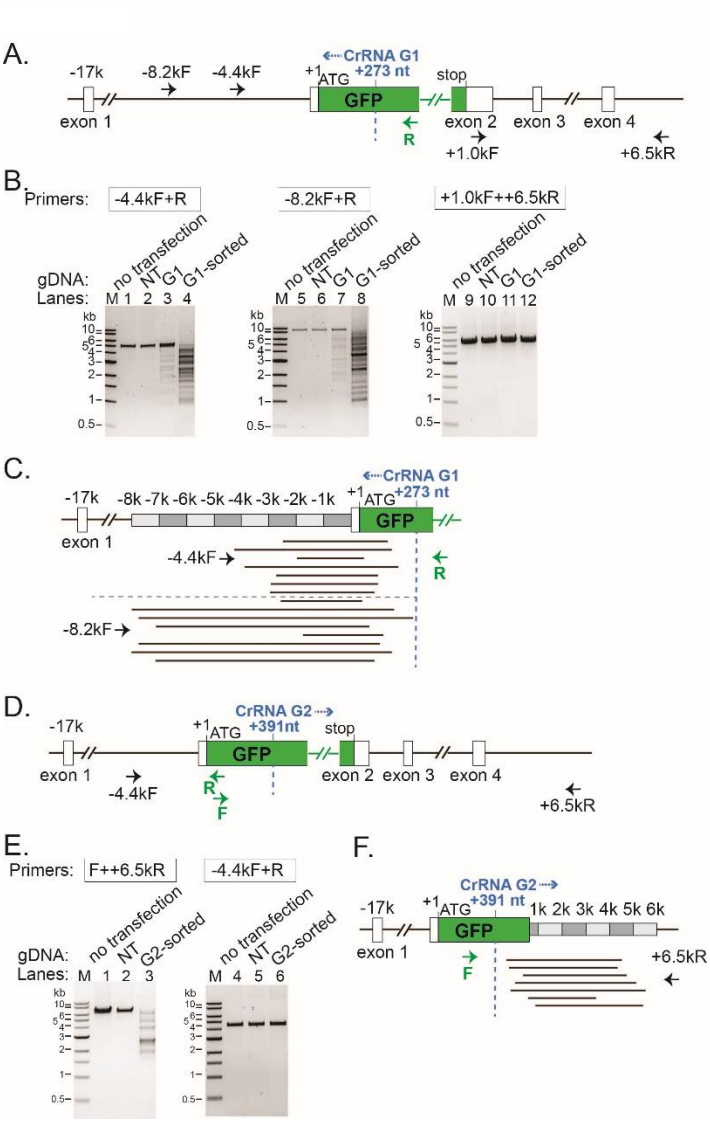


Figure 3.5: PCR and Sanger-sequencing based characterization of genomic lesions induced by Type I CRISPR-Cas. (**A**) Schematic of the EGFP reporter at DNMT3B locus and annealing sites for five PCR primers used in (**B**) and (**C**), and four primers used in (**E**) and (**F**). Positions relative to the EGFP translation start site (+1) are indicated. Recognition sites for Cascade G1 and G2 are marked by the dashed blue line. Blue arrowhead, direction of Cas3 translocation. (**B**) PCR-based genomic lesion characterization. A spectrum of chromosomal lesions was introduced upstream of EGFP by a EGFP-targeting Cascade-G1 and Cas3, in the sorted EGFP-negative population, as well as unsorted total cells. PCR primers used are indicated and their annealing sites depicted in (**A**). (**C**) Representative lesion locations revealed by cloning of the entire PCR in lanes 4 and 8 of (**B**) and Sanger sequencing. Black lines, deleted regions. (**E**) PCR-based lesion analysis. A collection of chromosomal deletions was induced downstream of EGFP by a EGFP-targeting Cascade-G2 and Cas3, in the sorted hESCs. PCR primers used are indicated and their annealing sites depicted in (**D**). (**F**). Representative lesion locations revealed by cloning and sanger sequencing of the PCR in lane 3 of (**E**). Black lines, deleted regions.

Table 3.1: Summary of Lesion Boundaries. 5’ and 3’ boundaries of lesions caused by GFP-targeting guides G1 or G2, identified via low throughput TOPO cloning and sequencing of GFP regions from FACS sorted GFP- cells. Positions listed are relative to the start (+1) of EGFP ORF.

Cascade-G1/Cas3 induced lesions, shown in Fig. 3.5C					
No #	5’ end	3’ end	No #	5’ end	3’ end
1	-2377	134	9	-7800	178
2	-4200	155	10	-7691	270
3	-2048	11	11	-7236	54
4	-3702	208	12	-1822	116
5	-2593	98	13	-7631	93
6	-2972	103	14	-7923	156
7	-3005	88	15	-7069	86
8	-2621	2			
Cascade-G2/Cas3 induced lesions, shown in Fig. 3.5F					
1	445	4478	5	635	4747
2	508	5432	6	648	3486
3	509	4429	7	802	5414
4	517	5046			

We further used a tagmentation- and next-generation sequencing (NGS)- based method to define Type I CRISPR-induced genome lesions more comprehensively. The genomic DNA of FACS-sorted, Cascade G1/Cas3 edited hESCs from the experiment in Fig. 1D was treated with adaptor-loaded Tn5 transposase, which randomly fragments DNA and attaches a single type of adaptor onto the fragmented ends. We then did a multi-step PCR using nested EGFP primers and a primer specific for the Tn5 adaptor to enrich for sequences spanning the junctions (**Fig. 3.6A**). The resulting NGS library was sequenced on an Illumina MiSeq, and 275 X 25 bp paired-end reads were analyzed to determine the extent of the corresponding deletions, as described in Methods. We detected lesion junctions in 8.9% of a total of 550,000 reads for the FACS-sorted sample, but in less than 0.01% of reads from the un-transfected control hESCs. Analysis of the lesion endpoints revealed that, consistent with Sanger sequencing results, the vast majority of 3' endpoints occur within a ~450 bp window upstream of the EGFP gene. The locations of the 5' boundaries are far more random and can occur tens of kilobases upstream (**Figs. 3.6B and 3.6C**). The majority of the Type I-E-induced chromosomal deletion sizes are distributed within a 10 kb range, however, a portion of deletions exists even up to 50 kb (**Fig. 3.6C**).

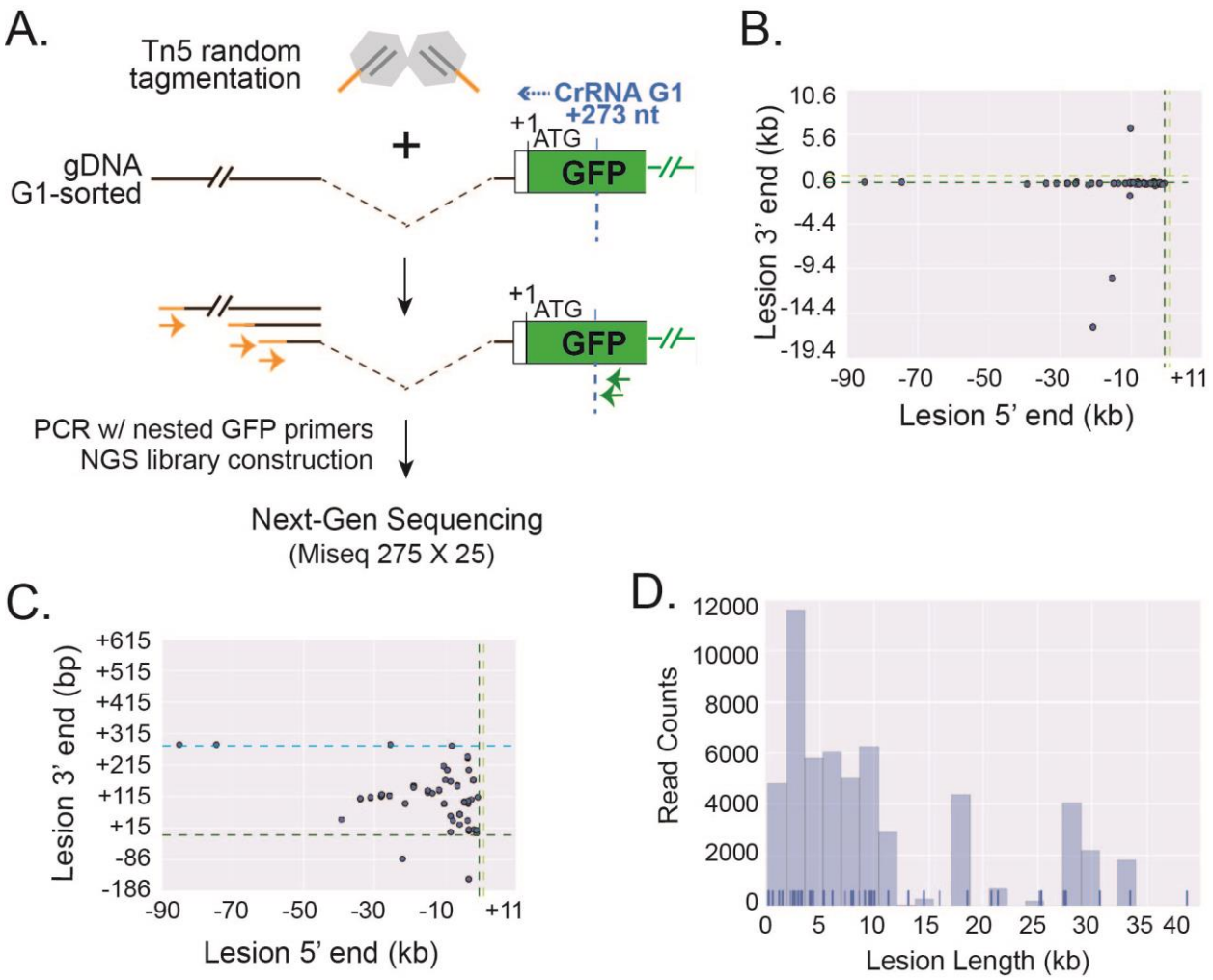


Figure 3.6: Tn5 and deep-sequencing based characterization of Type I CRISPR-induced genome lesions. (**A**) Schematics of Tn5 tagmentation-based NGS library construction. (**B**)-(**C**) Scatter plot for lesion-containing reads, showing the 5' (X-axis) and 3' (Y-axis) end points of chromosomal deletions, relative to the EGFP translation start site (+1) at the DNMT3B region. Dashed dark and light green lines, the start and end of the EGFP ORF; dashed blue line, recognition site for Cascade-G1. (**C**) is a zoom-in view of the same plot in (**B**). (**D**) Histogram showing the distribution of lesion lengths observed. Dark blue lines, precise locations of each observed lesion.

In its prokaryotic environment, Type I CRISPR interference typically eradicates targeted foreign DNA completely, or causes cell death if it were accidentally programmed against self-genome. The large chromosome size and strong intrinsic NHEJ activity attributable to human cells enabled us to explore the *T. fusca* type I-E CRISPR-Cas machinery at a much greater depth than previously described. Although the stochastic nature in the onset of genome deletion and the wide-range of deletion sizes was somewhat unexpected, this nonetheless opens new avenues for further mechanistic investigations. Moreover, an encouraging 13% genome editing efficiency could be achieved *ex vivo* in hESCs without the use of selectable markers. Also notable is that these results establish a lower bound on editing efficiency, not an upper bound, as disruption of the reporter will only be accomplished if the start of the lesion occurs in the first 270 bp for G1. Many more lesions may exist which were not detected by the reporter because the first double strand break of the lesion occurs outside of this range. The efficiency may be improved further by switching to a plasmid or mRNA-based delivery method to increase effector concentration and perdurance. Higher efficiencies may also be achieved by further improving *Tfu*Cascade's activity at 37 ˚C through directed evolution, or screening additional Type I CRISPR-Cas to identify more active gene-editors.

Various Cas9-based genetic screening methods now allow high-throughput functional genomics interrogation, which typically rely on a tiling library of sgRNA or sgRNA pairs (183, 198-201). The ability of the Type I system to introduce a spectrum of long-range deletions from a single target site would enable more effective and cost-effective non-coding CRISPR screens, because far fewer RNA guides are needed and each guide leads to a library of deletion mutants. Conceivably, precise removal of large disease-related genetic elements can also be achieved with a pair of Cascade/crRNAs spanning the pathological region. Moreover, Type I CRISPR can potentially be adopted for long-range epigenetic modifications, by accompanying Cascade with two different versions of Cas3, a small portion of helicase-dead Cas3 to initiate target DNA nicking, enabling the loading and translocation of nuclease-dead Cas3

fused with an epigenetic writer to introduce long-range epigenetic modifications. These Type-I CRISPR-based applications would greatly expand the genome engineering toolkit.

3.4 Acknowledgements

This work is supported by National Institutes of Health (NIH) grants GM118174 and GM102543 to A.K, GM117268 to Y.Z., and University of Michigan institutional fund and Biological Scholar Award to Y.Z. The authors thank Y. Xiao, J. Budhathoki, and R.A. Battaglia for helpful discussions, and Jaewon Heo for technical support.

3.5 Author contributions

A.D., Z.H, A.K. and Y.Z. designed the research. A.D. improved the activity of *T. fusca* Cascade and Cas3, adapted, programmed, and purified them for hESC editing experiments. Z.H. performed hESC editing experiments and deep-sequencings. M.J.G and Z.H. did PCR-based lesions analysis. S.E.H. created the dual-reporter hESC line, P.L.F. performed informatics analyses. A.D, Z.H., A.K., and Y.Z. wrote the manuscript with inputs from the rest of the authors. The authors declare competing financial interests. Correspondence and requests for materials should be addressed to A.K. (ak425@cornell.edu) or Y.Z. (yzhangbc@med.umich.edu).

3.6 Materials and methods

Information on plasmid expression plasmids and GFP/tdTomato plasmids is shown in **Table 3.2** Information on oligonucleotides used to generate those plasmids and for PCR analysis are shown in **Table 3.3**.

Expression and purification of TfuCas3 and TfuCascade

T. fusca Cascade and Cas3 was purified as described previously (57), with minor modifications. *Tfu*Cascade was recombinantly expressed in *E. coli* BL21cells in LB media using a three plasmid coexpression system. Cse1 is encoded in one plasmid with an N-terminal 6xHis-TwinStrep-SUMO tag

(pET19b). The rest of the Cascade components (Cse2, Cas7, Cas5e, and Cas6e) were encoded polycistronically in another vector (pCDF-Duet1), with a C-terminal NLS tag on Cas7. The crRNA was expressed from a synthetic CRISPR array containing three repeats and two spacers in ORF1 of pRSF-Duet1. Cells were grown at 37°C until the OD600 is between 0.6 and 1.0. Protein and RNA expression were induced by adding IPTG to a final concentration of 0.5 mM, and allowing the cell to grow overnight at 22˚C. 12 liters of cells were harvested and lysed by sonication in Lysis Buffer containing 30 mM HEPES pH 7.5 and 500 mM NaCl. The supernatant after centrifugation was loaded onto ~5 mL of StrepTactin resin and 2 mg per L of cells Avidin was supplemented to prevent cellular biotin from binding to the column. The column was washed with 3x15 ml of lysis buffer, and the protein was eluted with 10 ml of lysis buffer supplemented with 5 mM Desthiobiotin. After cleaving the TwinStrep-SUMO tag with SUMO protease overnight at 4 ˚C, *Tfu*Cascade was concentrated and buffer-exchanged to a buffer containing 30 mM HEPES pH 7.5 and 200 mM NaCl, and further purified on MonoQ. The final RNP was buffer-exchanged to 30 mM HEPES pH8.0 and 150 mM NaCl, sterilized with a syringe filter, concentrated to >20 μM, and flash-frozen for -80°C storage. To account for the nucleic acid component of *Tfu*Cascade, nanodrop UV 260/280 measurements were taken alongside a Bradford Assay standard curve. A conversion ratio was determined to more accurately estimate the concentration of the protein components.

*Tfu*Cas3 was expressed from M9 minimal media with an N-terminal TwinStrep-PreScission tag and a C-terminal 2xHA-NLS tag from a pET52b plasmid. A 5 ml starting culture was grown from LB media overnight at 37°C, propagated to a 100 mL M9 culture overnight at 37°C, then used to inoculate 3x2 L of M9 media. The trace metal supplement was left out of the standard M9 media to prevent Fe2+ incorporation into the Cas3 active site. 100 μM final concentration of cobalt chloride was added to the cell culture 30 minutes prior to IPTG induction, when the OD600 reached 0.6. Protein expression was induced by 1 mM IPTG overnight at 20°C. The cells were harvested, resuspended in lysis buffer (30 mM

HEPES pH 7.5 and 500 mM NaCl), lysed by sonication, and purified with a Strep-Tactin column similar to *Tfu*Cascade purification. The eluted protein was treated with PreScission protease overnight at 4°C to remove the TwinStrep tag. Cas3 was further purified over a HiLoad Superdex 200 size-exclusion column (SEC) equilibrated at 30 mM HEPES 7.5 and 150 mM NaCl. The main peak fractions were pooled and concentrated, flash-frozen in liquid nitrogen, and stored at -80°C until needed.

Table 3.2: Plasmids used in this study.

Plasmid names (Backbone/Insert)	Source
pET19b/TwinStrep-SUMO-wt CasA	Xiao *et al.*, 2017
pCDF-Duet1/N23A CasB, CasC-NLS, CasD, CasE	This paper
pCDF-Duet1/N23A CasB, CasC-NLS, rbs-CasD, CasE	This paper
pRSF/crRNA expression, streamlined	This paper
pRSF/crRNA-G-1 expression, streamlined	This paper
pRSF/crRNA-G-2 expression, streamlined	This paper
pRSF/crRNA-G-Td expression, streamlined	This paper
pRSF/crRNA-G-NT expression	Xiao *et al.* 2017
pET28b/Cas3-2xHA-NLS	This paper
hES-2A-DNMT3B-EGFP	This paper
hES-2A-DNMT3B-tdTomato	This paper

Cell Culture

Human ESCs were cultured in E8 medium on matrigel (Corning) coated tissue culture plates at 37˚C and 5% CO_2 in a humidified incubator, with daily media change. Cells were split every 4-5 days with 0.5 mM EDTA in 1x PBS.

Construction of hESC Dual-Reporter Line and DNMT3b targeting plasmids

Cells for transfection were harvested 2 days post passaging using TrypLE (Life Technologies) and resuspsended in OptiMem (Life Technologies) at a final concentration of 5 x 10^6 cells/mL. 500 µl of the cell suspension was added to a 0.4 cm cuvette containing 30 µg of the linearized DNMT3B-EGFP vector. Cells were electroporated using condition 320V, 200 µF, then plated on a 10 cm matrigel-coated dish in E8 media supplemented with 10 µM Y-27632 (Cayman Chemical). 0.5 µg/mL puromycin was added to the medium 3 days post-transfection and drug-resistant colonies exhibiting uniform EGFP expression were identified by fluorescent microscopy. A single EGFP+ clone was expanded and the puromycin selection cassette removed following electroporation of CRE recombinase mRNA. A subsequent round of targeting was performed as described above using the DNMT3B-tdTomato vector. Individual colonies expressing both tdTomato and EGFP reporters were identified, isolated and expanded. Successful biallelic targeting of the endogenous DNMT3B was confirmed by genomic DNA PCR using primers flanking the DNMT3B start codon.

To create DNMT3B targeting constructs (hES-2A-DNMT3B-EGFP and hES-2A-DNMT3B-tdTomato), a BAC clone (CTD- 2608L15) containing the complete DNMT3B coding region was obtained from the CalTech Human BAC Library (Life Technologies). Red-ET recombination was used to insert a DNA cassette encoding a tdTomato or EGFP reporter gene adjacent to a loxP-flanked PGK promoter driven puromycin resistance gene at the DNMT3B start codon in exon 2 of the DNMT3B gene. The ~40-kb SbfI fragment

containing the modified DNMT3B locus was then subcloned into the copy number inducible BAC vector, hES-2A. Prior to transfection these DNMT3B gene targeting vectors were linearized by SwaI.

RNP Electroporation of hESCs

The H9-DNMT3B-tdTomato/EGFP cells were electroporated using the Neon Transfection system (ThermoFisher) according to the manufacturer's instructions. Briefly, reporter cells were individualized with Accutase (ThermoFisher), washed once with DMEM/F12 (ThermoFisher) and resuspended in Neon buffer R to a concentration of $2x10^6$ cells/mL. 20-100 pmoles of NLS-TfuCascade and 20-60 pmoles of NLS-TfuCas3 were mixed with approximately 10^5 cells in buffer R in a 10 µL a total volume. This mixture was then electroporated with a 10 µL Neon tip (1100V, 20ms, 2 pulses) and plated in 24-well matrigel-coated plates containing 500 µL of E8 supplemented with 10µM Y-27632. The media was changed to regular E8 24 hours after electroporation. Cells were cultured in E8 with daily media change until analysis.

Flow Cytometry

Cells were individualized with Accutase 4-5 days after electroporation and resuspended in DMEM/F12 media immediately before experiments. For analysis, individualized cells were analyzed on an LSR Fortessa (BD) using 488nm laser for EGFP and 561nm laser for tdTomato. Data analysis was performed using FlowJo® v10.4.1. For FACS sorting, individualized cells were put on Synergy cell sorter (Sony) and GFP negative cells were sorted directly into a well of a 24-well plate coated with matrigel and filled with 1.5ml E8 media supplemented with 10 µM Y-27632 and 25 µg/mL recombinant human albumin (Sigma). Sorted cells were then cultured in tissue culture incubator with 5% CO_2 at 37˚C. Media was changed to regular E8 one day after sorting and daily media change with E8 was carried out thereafter, for 7-10 days.

Lesion Analysis by PCR

Genomic DNAs of hESCs were isolated using Gentra Puregene Cell Kit (Qiagen) per manufacturer protocol. PCRs in Fig. 3B were done using Q5 DNA Polymerase (NEB) with primer pair OYZ 438+478, OYZ 440+478 and HZG81+OYZ462, respectively. PCR products were resolved on 1% agarose gel stained by SYBR Safe (Invitrogen) and visualized with Chemidoc MP imager (Biorad). To characterize lesions shown in Fig. 3C, 40 µL of the lesion PCR reaction done for the sorted hECSc was purified using QIAquick PCR Purification Kit (Qiagen), and cloned into PCR-BluntII-TOPO vector (Invitrogen). Colony PCRs with M13 forward and reverse primers were carried out from the resulting colonies and large amplicons were Sanger sequenced. PCRs in Fig. 3E were perform with primer pair HZG707+OYZ462 and OYZ438+826. Lesions shown in Fig. 3F were characterized with TOPO cloning and Sanger sequencing.

Tn5 Tagmentation-based NGS Library Construction

Tn5 transposase was purified and loaded with one pre-annealed oligo pair ME-A/ME-rev as previously described (202). Tagmentation was performed in 10 mM Tris pH8.5, 5 mM $MgCl_2$, 8% PEG8000 using 60ng of genomic DNA and 700 ng of loaded Tn5, in a 20 µL total volume. After a 7 min incubation at 55˚C, tagmentation reactions were purified using QIAquick PCR Purification Kit. For NGS library construction, 1st step PCR amplification was carried out using Q5 DNA Polymerase for 15 cycles with oligos OYZ510+478, and then treated with Exonuclease I (NEB) to digest excess primers. 2nd step of nested-PCR was amplified for another 15 cycles with OYZ510+511. After Exonuclease I treatment, the 3rd step PCR was amplified for 10 cycles with OYZ510 and index primers. The final NGS libraries were purified using AMPure XP beads (Beckman Coulter, 1:1 ratio), eluted in 10 mM Tris pH 8.5, pooled together and sequenced on Illumina MiSeq with a Nano kit for 275 x 25 bp paired-end reads.

NGS Data Analysis

MiSeq sequencing reads were quality trimmed using Trimmomatic v0.33 (203) with filter settings "TRAILING:3 SLIDINGWINDOW:4:15 MINLEN:10", and then aligned to a defined window of the genome

spanning ~130 kb, which covered the entire DNMT3b locus with EGFP sequence inserted. Alignment was performed using bowtie2 v.2.1.0 (204) using default settings in local mode, allowing up to two alignments per read, and requiring stranded alignment of the forward read (originating from the Tn5 insertion end) with the forward direction of the reference sequence.

Subsequently, we filtered the alignments to identify potential chromosomal junction events as follows: a single read with two alignments was considered a junction if one alignment had cigar string *w*M*x*S and the other alignment had cigar string *y*S*z*M (corresponding to a case where one alignment matched the first portion of the read, and the other matched the second portion of the read); we further required that the magnitude of the difference between *w* and *y* be no greater than 10. Junctions were then assumed to occur between the position in the reference sequence *y* bp to the right of the leftmost end of the left alignment, and the position at which the unclipped portion of the right alignment begins. Count tables were obtained indicating the junctions corresponding to each lesion event. *N.b.* most lesions were observed repeatedly across many reads.

QUANTIFICATION AND STATISTICAL ANALYSIS

All genome editing and *in vitro* experiments were repeated three times, representative FACS plots and gel images were shown.

Table 3.3. Oligonucleotides used in this study.

No.	Oligo Name	Sequence, 5'-3'	Purpose
For Cascade, Cas3 expression and purification.			
0117	CasC NLS 2 Fwd	GGCAAGCTTCCCAAGAAGAAGAGGAAGGTGGAGGGGGAGCGGGAG TGAGTG	pCDF-Duet1 / wt CasB, CasC-NLS, CasD, CasE
0118	CasC GSLink 2 Rev	ACCTGAACCGCTACCGAAGGCTGCCGCGACCATGG	
0136	CasD RBS Fwd	TTTAAGAAGGAGATATACATATGAGTGGCTTCCTGCTGCGGCTA	pCDF-Duet1 / wt CasB, CasC-NLS, rbs-CasD, CasE
0137	CasD RBS Rev	attaaagttaaacaaaaTTATTCCCGCTCCCCCTCCACC	
0134	pRSF conversion for crRNA Fwd	tcatcgaatttttgcagcag	pRSF/crRNA expression, streamlined
0127	pRSF conversion for crRNA Rev	accatggcctatagtgagtcgtattaatttcctaatgc	
0113	Cas3 CTD+2XHA-NLS (gBlock)	cacgtgctcgcgacccggttcggtgccggttcagtccgggtgttgtgctactacgtggacacg gcggggaaccgctggcttgaccctgaatgcacggtcgagtttcctgaacagggcacggggc gagagggccggttcaccatggcagactgccgcgacctggtggcccgcacgatcccggtgcg tatgggtccctgggcgagtcaactcaccgaggacaaccatcctcctgaggcatggcgggag tcgttctaccttcgcgacctggttcttatacctcaacgtgtgacagacgagggcgcggtgctc cccactgaaaccggtggacgagagtggttgcttgatccctgtaaggggctgatctttGGAT CCGTTggtTACCCATACGATGTTCCTGACTATGCGGGCTATCCCTATGA CGTCCCGGACTATGCAGGATCCTATCCAGAATTCcccaagaagaagaggaa ggtgtAactcgag	pET28b/Cas3-2xHA-NLS
	GFP crRNA-G1 Synthesis	gaattcGAGCCCCACGCACGTGGGGATGGACCGGCTACGTCCAGGAGC GCACCATCTTCTTCAAGGTGAGCCCCACGCACGTGGGGATGGACCGG	pRSF/crRNA-G-1

		CTACGTCCAGGAGCGCACCATCTTCTTCAAGGTGAGCCCCACGCACGT GGGGATGGTGACaagctt	
	GFP crRNA-G2 Synthesis	aagcttGAGCCCCACGCACGTGGGGATGGACCGTCGATGCCCTTCAGC TCGATGCGGTTCACCAGGTGAGCCCCACGCACGTGGGGATGGACCGT CGATGCCCTTCAGCTCGATGCGGTTCACCAGGTGAGCCCCACGCACG TGGGGATGGTGACgcggccgc	pRSF/crRNA-G-2
	crRNA-NT Synthesis	gaattcTAATACGACTCACTATAGGGAGCCCCACGCACGTGGGGATGG ACCGCCAGTGATAAGTGGAATGCCATGTGGGCTGTCGTGAGCCCCAC GCACGTGGGGATGGACCGCCAGTGATAAGTGGAATGCCATGTGGGC TGTCGTGAGCCCCACGCACGTGGGGATGGACCGCCAGTGATAAGTG GAATGCCATGTGGGCTGTCGTGAGCCCCACGCACGTGGGGATGGAC CGCCAGTGATAAGTGGAATGCCATGTGGGCTGTCGTGAGCCCCACGC ACGTGGGGATGGACCGCCAGTGATAAGTGGAATGCCATGTGGGCTG TCGTGAGCCCCACGCACGTGGGGATGGACCGCTAGCATAACCCCTTG GGGCCTCTAAACGGGTCTTGAGGGGTTTTTTggatcc	pRSF/crRNA-NT
	tdTomato crRNA synthesis	ATATCATAGTACAATAGGATCCGAGCCCCACGCACGTGGGGATGGAC CGctggacatcacctcccacaacgaggactacacGTGAGCCCCACGCACGTGGG GATGGACCGGAATTCAGTCGTAGTTTCGCGCATCATGGCCATA	pRSF/crRNA-G-Td
For PCR-based lesion analysis			
OYZ438	DNMT3b-F-4.4k	GATGGGGTGGGGGTTAAAGG	PCR, Fig. 3B
OYZ440	DNMT3b-F-8.2k	AGTACTGCACTCTTTGCCCC	
OYZ478	EGFP-Rev	acgaactccagcaggacc	PCR, Fig. 3B; NGS library
HZG81	DNMT3b-after EGFP-Fwd (1.0k)	aagggagacaccaggcatc	PCR, Fig. 3B

HZG531	DNMT3b-exon3 Rev	gagagtcgcgagcttgatct	
oYZ826	EGFP-62-81-Rev	CTTGTGGCCGTTTACGTCGC	PCR, Fig. 3E
HZG707	EGFP-fwd	gtgagcaagggcgaggag	
oYZ462	DNMT3b-R-6.5k	GGCCAATTACTGGGTTCAGG	
For Tagmentation and NGS library preparation			
OYZ507	Tn5-ME-rev	/5phos/CTGTCTCTTATACACATCT	Tn5 loading
OYZ508	Tn5-ME-A	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAG	
OYZ510	Nextera P5	AATGATACGGCGACCACCGAGATCTACACTCGTCGGCAGCGTC	NGS library prep
OYZ511	EGFP-302-Illumina	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTggtcttgtagttgccgtcg t	
N/A	Index 1	CAAGCAGAAGACGGCATACGAGATGAAAGACCGTGACTGGAGTTCA GACGTGTGCT	
N/A	Index 2	CAAGCAGAAGACGGCATACGAGATATGGGCACGTGACTGGAGTTCA GACGTGTGCT	

CHAPTER IV – SPECULATION AND CAS3 DIMERIZATION

4.1 Type I Genome Editing Future Directions

Chapter III of this thesis describes a novel genome editing platform based on the Type I-E CRISPR system. Other Type I-E systems exist, many of which have evolved at a temperature compatible with mammalian genome editing. Screening of other Type I-E systems might reveal a system which is as easy to work with as *T. fusca*, but with a more effective R-loop behavior at editing temperatures. Alternatively, other Type I systems exist which contain fewer unique subunit proteins, such as the Type I-C system of bacteria such as *Bacillus halodurans* (59). These might be more effective in a plasmid delivery system due to the fact that the overall coding sequence is shorter.

The wild-type *T. fusca* Cascade complex should also be tested in genome editing. The trade-off observed in EMSA results might not represent the optimal situation. Despite Wild-Type Cascade having a higher affinity for non-specific interactions, the mutation of N23A might come with other costs. Since the limiting factor appears to be target binding, evidenced by the fact that Cascade concentration manipulation alters editing efficiency much more than Cas3 concentration manipulation, it would be prudent to more thoroughly screen other mutations as well as to test the Wild-Type in genome editing.

4.2 Acquisition Speculation

Spacer biogenesis remains one of the largest unanswered questions in the field of CRISPR biology. Both for naïve and primed acquisition, the precise mechanism that provides the substrates for Cas1/Cas2 to integrate spacers into the genome are not well understood. There is, however, work that has characterized what Cas1/Cas2 spacer substrates presumably look like *in-vivo*. These substrates have been shown to likely possess at least one forked end in structures of Cas1/Cas2 bound to substrate DNA solved in our lab and by others (23, 205). Single-molecule work in our lab has also shown that

processing can occur after the Cas1/Cas2 binding event in the presence of host exonucleases (unpublished). This means that the forked end can be inserted to the CRISPR array and the other side can be processed afterwards. This can represent an additional mechanism for selecting the orientation of spacer insertion in the CRISPR leader-repeat region, since the same work has shown that a substrate with two forked ends seems to insert bi-directionally. The sequence of events in this model would be that Cas1/Cas2 binds a single-end forked DNA substrate that has been produced in an undefined process which ensures a PAM site at the forked end. Then, integration at the leader side of the CRISPR locus is catalyzed by Cas1/Cas2. Next, host exonucleases chew the un-integrated side of the spacer, stopping where Cas1/Cas2 protects the substrate. Finally, spacer-side integration follows. But this model still cannot resolve the resulting post-integration structure. There are several hypotheses that involve different sets of repair machinery – since the CRISPR locus is constantly transcribed, a plausible theory is that transcription-mediated repair is involved in the resolution of this integration intermediary.

It has been shown that in the Type II system, the Cas9 protein is required for PAM specificity in acquisition (94). Further, when the PAM specificity of Cas9 is changed by substituting a closely related orthologue, it results in a change in acquisition specificity for targets adjacent to the substituted Cas9's PAM preference. This shows that PAM recognition for spacer acquisition relies on the interference module in the Type II system. I believe this strongly hints towards a role for Cascade or Cas3 in spacer acquisition in the Type I system generally, and in selection of PAM-containing sequences specifically.

4.3 Dimerization

The mechanism by which Cas3 induces the double strand breaks observed in editing experiments is currently not well understood. Even though cleavage has been observed on both strands of DNA in some experiments, others observe obligate single-stranded activity (80, 82, 158). What causes this discrepancy? One possible explanation is that Cas3 biochemistry involves multiple modes of activity.

One mode would be the well-characterized reeling behavior observed in single-molecule DNA curtain experiments (81-83). In these setups, double strand breaks have not typically been reported in the literature. In the appendix of this thesis, the first description of Cas3-mediated double strand breaks observed via DNA curtain is reported. However, it is still possible that double strand breaks are caused in multiple ways both by dimerization and by a single Cas3, with the roadblock-dependent mechanism relying on either tension caused by Cas3's helicase activity to induce the break or by Cas3's nuclease domain having a rare interaction with the opposite strand of DNA. Further, the experimental design is unable to distinguish between Cas3 dimerization and double strand breaks caused by a single Cas3, as the Cas3 used in these experiments is not fluorescently labeled. The description of the double strand break and Cas3 roadblock experiments was not intended to rule out the possibility that the observations may have been made when excess Cas3 present in solution, allowing for dimerization to occur (see Materials and Methods of Appendix). This conclusion is supported by the fact that double strand breaks were still observed in the absence of a roadblock to Cas3 translocation and the break sites were spread out over most of the substrate (Figure AI.5c).

Further evidence for the existence of a Cas3 dimer is that it has already been observed. When the *T. fusca* Cas3 crystal structure was originally solved, the unit cell contained four Cas3 monomers arranged as two dimers (171). The arrangement of the individual dimers was head-to-tail, with the nuclease of one monomer next to the helicase of the other. The interface between the two Cas3s which comprised the dimer could be interpreted as more extensive than would be expected from a biologically irrelevant crystal packing interface (Figure 4.1a). With the more recent information from the high-resolution Cryo-EM structure, the interface surfaces can be assigned as the same ones responsible for Cascade-Cas3 interaction (57). If the Cas3 dimer observed in the Huo *et al.* structure is biologically relevant, then it can safely be concluded that the Cas3 dimer activity is mutually exclusive to the Cascade-Cas3 activity (194).

Cas3's combination nuclease-helicase behavior has been compared with RecBCD (85). Both are helicase-nuclease fusions, but RecBCD has bi-directional helicase behavior (97). RecBCD acts as a component in the innate bacterial immune system with the purpose of degrading exposed DNA processively. However, RecBCD has another function as a detector of Chi sites and an initiator of recombination (97). Cas3 as a monomer has only 3'-5' helicase activity, however if it were present in a dimer, it is possible that it could travel bi-directionally. Further parallels between RecBCD and Cas3 may exist in the detection of specific motifs in the substrate DNA – PAMs for Cas3 instead of RecBCD's Chi sites. This would result in a natural precursor for spacer biogenesis in CRISPR adaptation.

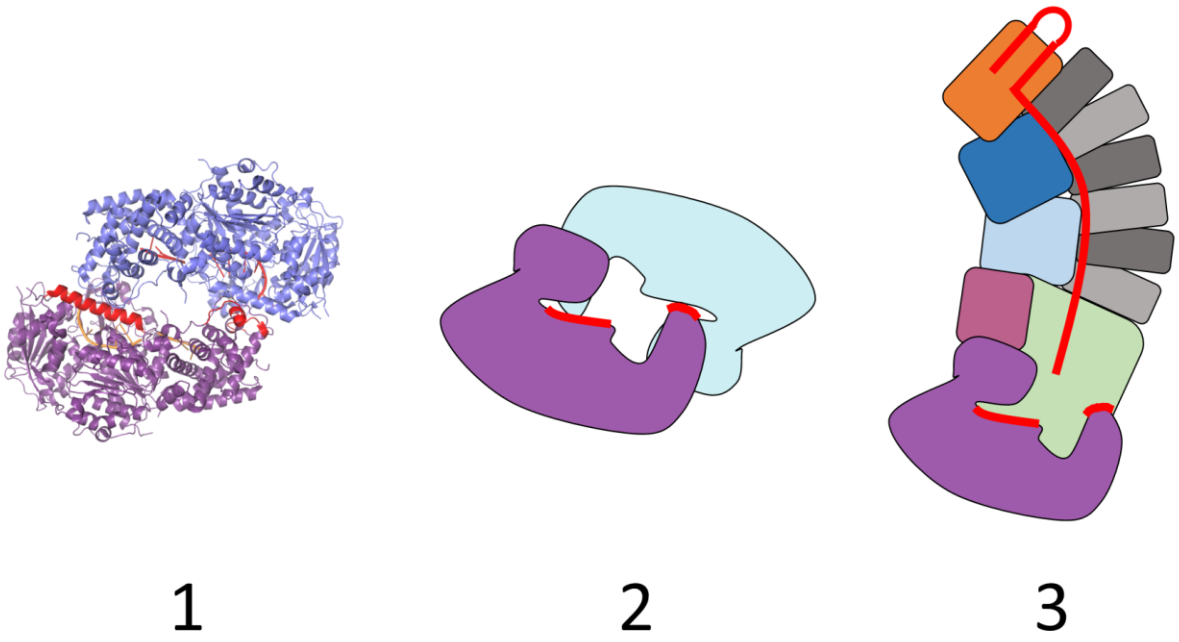
Towards this end, I have begun work to test the hypothesis that the Cas3 dimer observed in the crystal structure might be biologically relevant. A framework for studying Cas3 dimerization plausibly involves observing the transition of Cas3 from Cascade-bound DNA reeling activity to Cascade independent translocation. This suggestion is strengthened by the fact that in hESC cells, Cas3-induced double strand breaks occur predominantly distal from the target site and only occur in the direction that Cas3 is expected to move, suggesting that Cas3's helicase must be active for some time before DSB's are induced (Chapter III). This observation, combined with the observation that Cas3 has two modes of activity (Cascade-bound reeling vs. Cascade-independent translocation), naturally leads to the hypothesis that this regional transition relative to the Cascade target site from DSB-prohibitive to DSB-permissive might coincide with a transition in Cas3's behavior.

Structural investigation of the question is aided by the fact that we already possess a putative structure of the complex at high resolution, allowing the identification of residues involved in the dimerization surface. However, complicating the problem is the fact that the interface shares many contacts with the Cascade-Cas3 interface. This makes functional assessment of dimerization more difficult. Cas3 activity independent of Cascade has been difficult to observe, and since Cas3 alone has very little or no activity

on double-stranded DNA, functional studies of Cas3 dimerization interface perturbation would likely have to occur at sites which are independent of the Cascade-Cas3 interaction surface.

I decided to test the Cas3 dimerization interface through mutational analysis anyway, just in case the Cascade-Cas3 interface could be separated from the Cas3 dimerization interface. I generated several mutants at positions S171R, L794E, and A790E. When Wild-Type Cas3 is subjected to Size Exclusion Chromatography, a small but reproducible dimer peak has been observed (Figure 4.1b). When these Dimer interface mutants were purified over SEC, some showed a decreased or absent dimer peak. This suggests that perturbation of the putative interface is having some effect, though specific conclusions cannot be drawn from this data alone. When these mutants were tested in a Cascade-mediated substrate cleavage assay, unambiguous interpretation of the results were difficult (Figure 4.2). The S171R mutant appeared to have activity comparable to Wild-Type on both the target and non-target strands of the substrate. A790R, however, showed an overall decrease in substrate cleavage, but interestingly an increased bias towards generating nicks on the Target strand at the 3' end.

a)



b)

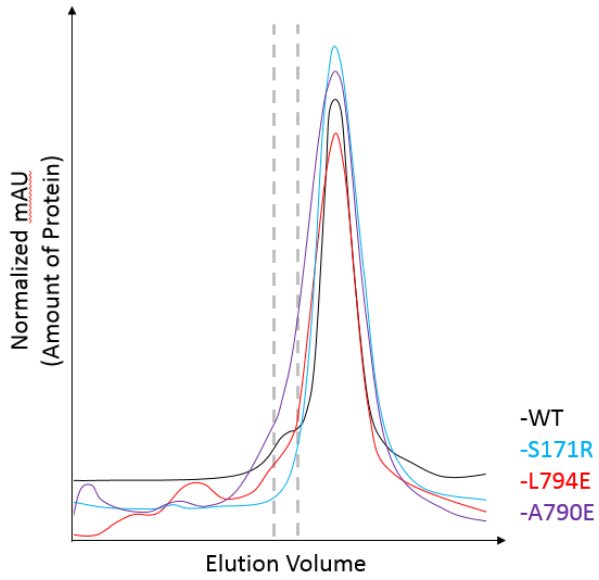
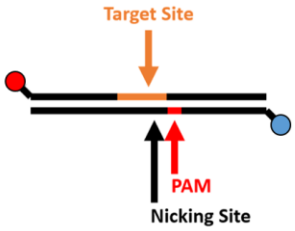
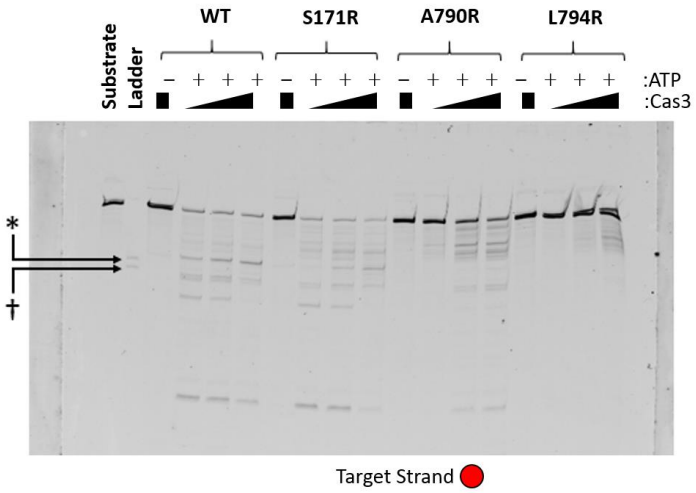


Figure 4.1. **Cas3 dimerization structure and initial mutational analysis** a) 1. Cas3 Dimer Crystal Structure. The Cas3 dimer interface is highlighted in red on the purple sub-unit. 2. Cartoon representation of the dimer. 3. Cascade/Cas3 interface cartoon. b) Cas3 dimer mutants SEC profiles. WT Cas3 exhibits a clear dimer shoulder while dimer mutants exhibit a decreased or non-existent dimer fraction.

a)



b)



c)

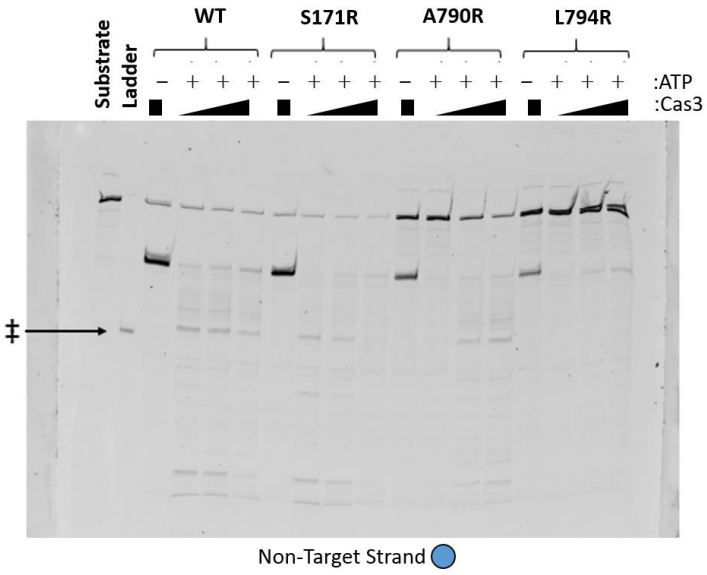
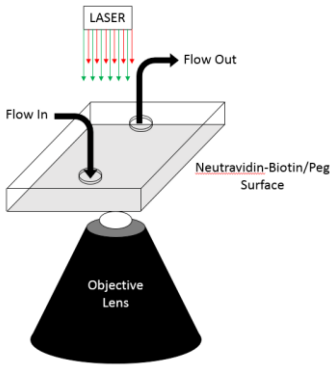


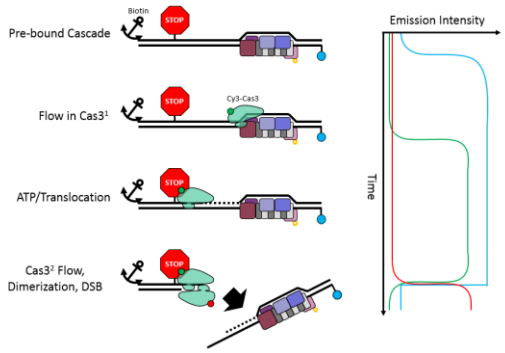
Figure 4.2 Cas3 dimer mutants activity assay a) Fluorescent substrate used in activity assay showing in b and c. b) Activity assay showing the relative nuclease activity of wildtype Cas3, S171R, A790R, and L794R acting on the target strand of the fluorescent substrate in a. c) The same gel from b, scanned for activity on the non-target strand. *NTS-Nick Site | †NTS-R-loop Start | ‡Nick-50bp Cas3 concentrations are 50 nM, 200 nM, 800 nM.

It became apparent that bulk biochemical investigation of the dimerization hypothesis would prove to be exceptionally difficult since controlling the formation of the dimer would be difficult in reaction conditions where Cas3 would always necessarily be provided in excess. Performing the experiment in a flow-cell allows the substrate to be bound to the surface and proteins and reagents can be flowed in sequentially in a controlled manner (Figure 4.3a). This allows for Cascade bound to substrate on the surface of the flow cell to recruit a single Cas3 in the absence of ATP. Excess Cas3 can be washed away, leaving the Cascade-bound Cas3 stably associated. Addition of ATP initiates Cas3 translocation, and subsequent addition of Cas3 theoretically allows for more control over the formation of the putative Cas3 dimer. The Cas3 dimer model also lends itself to single-molecule FRET (smFRET) investigation. If two populations of Cas3 labeled with a FRET-compatible fluorophore pair were used in smFRET experiments, dimerization would have an unmistakable and clear signal that proves the formation of the dimer on substrate DNA (Figure 4.3b).

a)



b)



c)

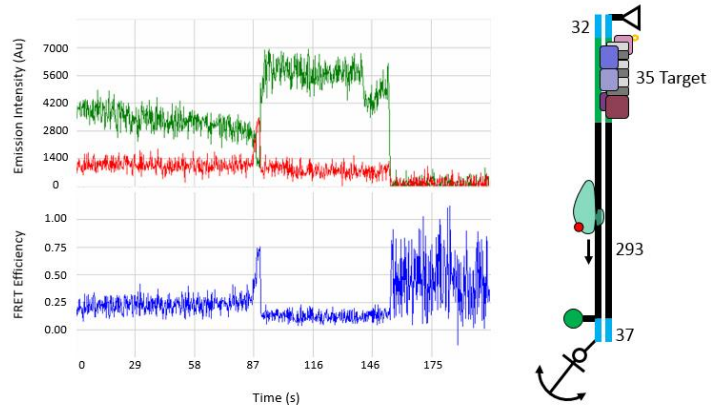


Figure 4.3: Framework for testing Cas3 Dimerization by smFRET. a) Schematic of smFRET flow cell. **b**) Design of ideal smFRET Cas3 dimerization experiment. The first fluorophore-labeled Cas3 is loaded to Cascade and synchronized at an impassible roadblock by providing ATP. Then, a second population of fluorophore-labeled Cas3 is flowed in, FRET signal is observed and a loss of the third fluorophore confirms the DSB. **c**) FRET trace of Cas3 activity in the flow-cell, showing Cas3 translocation up to and past the C6amino-dT modified with NHS-Cy3.

As initial tests of the *T. fusca* Cascade/Cas3 system in single-molecule experiments, I prepared a substrate that contains a C6-amino-dT modification which was labeled with NHS-Cy3. This substrate also contained a Cascade target site such that Cas3, after being recruited to the substrate by Cascade, would track along the strand and approach the fluorophore on the same strand. In previous helicase characterization studies, C6-amino-dT-fluorophore has had differing conclusions on whether or not the modification was capable of stalling various helicases. When Cy5-labeled Cas3 approached the fluorophore, smFRET traces showed that it is capable of overcoming, bypassing, or dissociating from the obstacle (Figure 4.3c). Some other traces showed that Cas3 might pause in some fraction of these events, but not for very long. These experiments served as an encouraging sign that the system works well and in a predictable manner, but would not ultimately help in the determination of whether or not dimerization occurs.

Towards this end, a new substrate with an *E. coli* Ter site near the biotin-labeled terminus was designed with a C6-amino-dT modification. The Ter site is a strong DNA roadblock used during *E. coli* genomic replication and is bound by the protein Tus. The Ter site is a direction-dependent roadblock to replication such that it only prevents translocation along DNA from a single side, and so the roadblock site was designed such that the non-permissive orientation was presented to Cas3 translocation. This experimental setup with Tus as a roadblock to Cas3 translocation remains to be tested.

APPENDIX I

APPENDIX I CREDITS:

This work constitutes a collaboration with the Finkelstein lab at University of Texas and was published in Cell in 2018. As described in the author contributions section Maxwell Brown, Kaylee Dillard, Logan Myler, Yibei Xiao, Ailong Ke, and Ilya Finkelstein conceived of the study. The DNA curtain experiments were performed by Maxwell Brown, Kaylee Dillard, and Logan R. Myler. Yibei Xiao and I purified protein used in the study, with my primary contribution being the provision of Cobalt-purified Cas3, which increased the rate that Cas3 activity was observed, improving the quality of the study. I also performed the in-vivo experiments to assay mutant Cascade complexes. Erik Hernandez and Samuel Dahlhauser provided fluorescent peptides. Yoori Kim provided software used for data analysis. The manuscript was written by Maxwell Brown, Kaylee Dillard, and Ilya Finkelstein.

Assembly and translocation of a CRISPR-Cas primed acquisition complex

Maxwell W. Brown,[1,¶,*] Kaylee E. Dillard,[1,¶,*] Yibei Xiao,[2] Adam Dolan,[2] Erik Hernandez,[3] Samuel Dahlhauser,[3] Yoori Kim,[1] Logan R. Myler,[1] Eric Anslyn,[3] Ailong Ke,[2] and Ilya J. Finkelstein[1,4,*]

[1] Department of Molecular Biosciences and Institute for Cellular and Molecular Biology, University of Texas at Austin, Austin, Texas 78712, USA

[2] Department of Molecular Biology and Genetics, Cornell University, 253 Biotechnology Building, Ithaca, NY 14853, USA

[3] Department of Chemistry, University of Texas at Austin, Austin, Texas 78712, USA

[4] Center for Systems and Synthetic Biology, University of Texas at Austin, Austin, Texas 78712, USA

* Corresponding authors: maxwellbrown@utexas.edu, kaylee.dillard@utexas.edu, ifinkelstein@cm.utexas.edu

¶ These authors contributed equally to this work

Keywords: CRISPR, Cascade, Primed Acquisition, DNA Curtains, Fluorescence Microscopy

AI.1 Abstract

CRISPR-Cas systems confer an adaptive immunity against viruses. Following viral injection, Cas1-Cas2 integrates segments of the viral genome (spacers) into the CRISPR locus. In type I CRISPR-Cas systems, efficient “primed” spacer acquisition and viral degradation (interference) require both the Cascade complex and the Cas3 helicase/nuclease. Here, we present single-molecule characterization of the *Thermobifida fusca* (*Tfu*) primed acquisition complex (PAC). We show that *Tfu*Cascade rapidly samples non-specific DNA via facilitated one-dimensional diffusion. Cas3 loads at target-bound Cascade and the Cascade/Cas3 complex translocates via a looped DNA intermediate. Cascade/Cas3 complexes stall at diverse protein roadblocks. In contrast, Cas1-Cas2 samples DNA transiently via 3D collisions. Moreover, Cas1-Cas2 associates with Cascade and translocates with Cascade/Cas3, forming the PAC. PACs can displace different protein roadblocks, suggesting a mechanism for long-range spacer acquisition. This work provides a molecular basis for the coordinated steps in CRISPR-based adaptive immunity.

AI.2 Introduction

Bacteria and archaea destroy foreign nucleic acids by mounting an RNA-based CRISPR adaptive immune response[1–3]. In Type I CRISPRs, the most frequently found CRISPR sub-type in bacteria and archaea[3,4], foreign DNAs that trigger efficient immunity can also provoke primed acquisition of protospacers into the CRISPR locus[5–12]. Both interference and primed acquisition require Cascade (CRISPR-associated complex for antiviral defense) and the Cas3 helicase/nuclease. Primed acquisition also requires the Cas1-Cas2 integrase, however the biophysical mechanisms of how interference and primed acquisition coordinate have remained elusive. Here, we present single-molecule characterization of the Type I-E *Thermobifida fusca (Tfu)* primed acquisition complex (PAC). *Tfu*Cascade rapidly samples non-specific DNA for its target via facilitated one-dimensional diffusion. An evolutionary-conserved positive patch on the Cse1 subunit promotes facilitated diffusion and increases the target recognition efficiency. Conformational locks stabilize the complex during R-loop propagation, even on partially complementary DNA targets. Cas3 loads at target-bound Cascade and the Cascade/Cas3 complex initiates processive translocation via a looped DNA intermediate. Moving Cascade/Cas3 complexes stall and release the DNA loop at protein roadblocks. Cas1-Cas2 samples DNA transiently via 3D collisions, but is stabilized via protein interactions with target-bound Cascade. Cas1-Cas2 also remains associated with Cascade/Cas3 and is further stabilized as part of the translocating PAC. By directly imaging all key sub-complexes involved in target recognition, interference, and primed acquisition, this work provides a molecular basis for the central steps in CRISPR-based adaptive immunity.

CRISPR adaptive immunity consists of three main activities: interference, primed acquisition, and naïve acquisition[13–16]. Interference provides immunity by targeting and destroying foreign nucleic acids that are recorded in the CRISPR locus. The CRISPR system adapts to new threats by acquiring and segments of foreign genetic elements into the CRISPR array, where they are transcribed and used to confer immunity against the invading nucleic acid[17–20]. In Type I CRISPRs, the Cascade surveillance complex initiates both interference and primed acquisition[5–12]. Cascade surveils the cell for foreign DNA that is complementary to its CRISPR RNA (crRNA)[17]. An RNA-DNA loop (R-loop) between the crRNA and the duplex target DNA conformationally locks Cascade onto the foreign genetic element[21–29]. Next, target-bound Cascade loads Cas3 nuclease/helicase, which unwinds and degrades the foreign DNA[21,30–33].

Primed and naïve acquisition both require the Cas1-Cas2 integrase. Cas1-Cas2 inserts new protospacers into the CRISPR locus in the host's genome via a cut-and-paste transposase mechanism[34–36]. Naïve acquisition can integrate foreign nucleic acids that the cell has not encountered previously and requires host nucleases to produce substrates for Cas1-Cas2[37]. In contrast, primed acquisition uses Cascade/Cas3 to produce protospacers that Cas1-Cas2 then integrates into the CRISPR locus[5–11]. Primed acquisition thus requires a prior record of infection by a related pathogen. Because primed acquisition is substantially more efficient than naïve acquisition, this mechanism permits the cell to rapidly adapt to phages that have acquired escape mutations[5,9,11,38]. Although the genetic and biochemical basis for primed acquisition have been established previously, the biophysical mechanisms underpinning interactions between Cascade, Cas3, and Cas1-Cas2 have remained elusive[5,6,9]. To address this gap, we

report the stepwise assembly and biophysical characterization of the *Thermobifida fusca (Tfu)* Type I-E CRISPR-Cas interference and primed acquisition machineries. Using single-molecule fluorescence imaging of each sub-complex, we show that Cse1 plays a key role in target recognition by promoting rapid scanning of foreign DNA via facilitated diffusion. After target recognition, Cascade recruits Cas3, and the Cascade/Cas3 interference sub-complex translocates via a looped DNA intermediate. Finally, we provide direct evidence that Cascade/Cas3 interacts with Cas1-Cas2 to form a translocating complex that combines all the biochemical functions required for both interference and primed acquisition.

AI.3 Cse1 promotes target recognition via facilitated diffusion on non-specific DNA

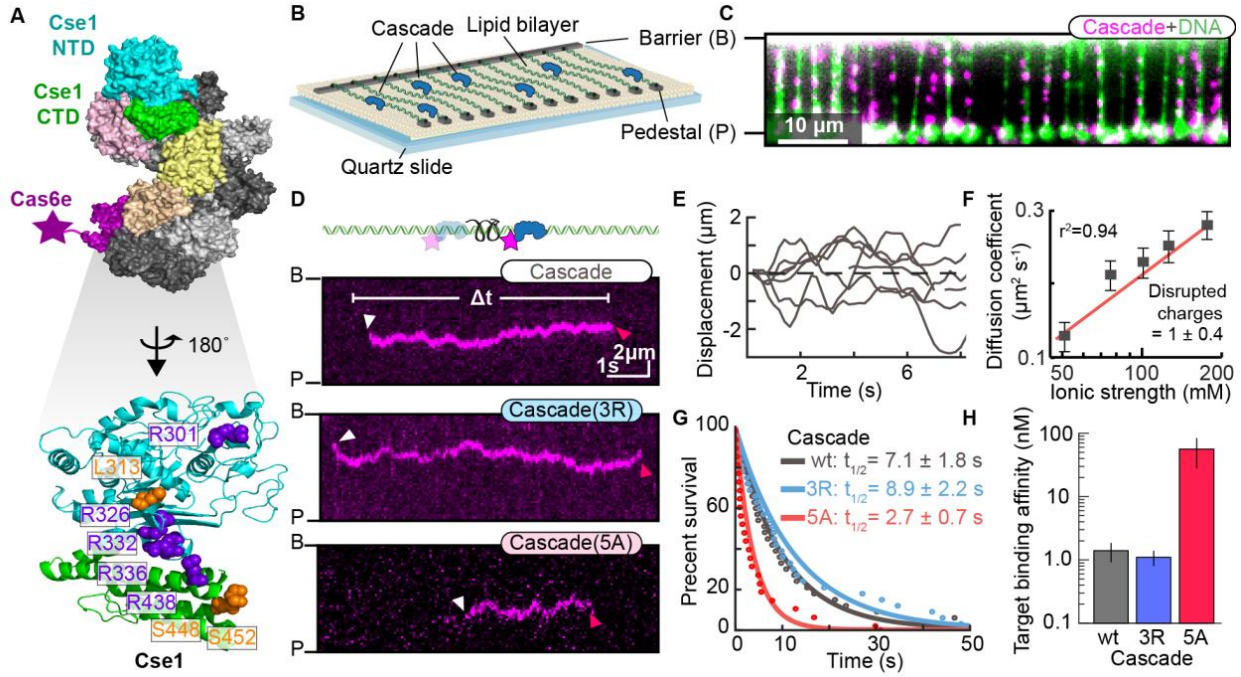
To understand how Cascade participates in both interference and primed acquisition, we first imaged fluorescent *Tfu*Cascade on double-tethered DNA curtains that extend the substrate in the absence of buffer flow[39,40] (**Figures AI.1 and S1**). The DNA substrate (derived from bacteriophage λ) lacked a target DNA that was complementary to the Cascade crRNA. Prior studies reported that the *S. pyogenes* Cas9 and *E. coli* Type I-E effector complexes sample protospacer adjacent motif (PAM) sites exclusively via three dimensional (3D) collisions, suggesting that this is a universal feature of diverse CRISPR systems[41,42]. Unexpectedly, 90% (N=258 out of 288) of *Tfu*Cascade molecules initially bound non-specific DNA and scanned the substrate via facilitated one-dimensional (1D) diffusion (**Figure AI.1D-F**). Facilitated diffusion can accelerate the target search dynamics, as has been observed for other DNA-binding proteins[43]. During 1D diffusion, proteins can either slide along the helical pitch of the DNA

backbone, or can transiently dissociate and associate with the DNA via a series of microscopic hops. Hopping allows proteins to efficiently search larger segments of the genome while frequently randomizing the spatial register between the protein and the DNA backbone (see below)[43,44]. Hopping can be observed indirectly by measuring the change in the diffusion coefficients at higher ionic strengths, which increase electrostatic screening between the protein and DNA. This results in measurably larger 1D diffusion coefficients and can be used to estimate the number of disrupted electrostatic charges[45]. Cascade diffusion coefficients increased with higher ionic strength, with approximately one charge screened at physiological ionic strength (**Figure AI.1F**). Cascade lacking Cse1 did not diffuse on DNA curtains. Therefore, we conjectured that a positive patch on the *Tfu*Cse1 outer surface (**Figure AI.1A, bottom**) promotes facilitated diffusion of Cascade during foreign DNA surveillance[46]. A structure-based multi-sequence alignment of divergent Cse1 variants revealed that the positive patch is highly conserved and can extend up to eight amino acids (**Figure S2A**)[47]. Notably, this positive patch is disrupted in the *E. coli (Ec)* Cse1, likely limiting the 1D scanning mode of *Ec*Cascade beyond the resolution of prior studies (**Figure S2B**)[27,48–50]. The *Tfu*Cse1 studied here encodes positive charges at five of these eight sites (**Figure AI.1A**). To test the importance of the Cse1 positive patch on facilitated diffusion, we purified Cascade harboring Cse1(5A), a variant with all five positive residues mutated to alanine (**Figure S2B**). Cse1(5A)-Cascade diffusion trajectories were 2.6-fold shorter than the wild type complex on non-specific DNA (**Figure AI.1G**; 2.7 ± 0.7 sec, N=50 molecules vs. 7.1 ± 1.8 sec, N=100), and also had a 50-fold lower binding affinity for target DNA, as determined by electrophoretic mobility shift assays (EMSAs, **Figures AI.1H and S2**). Extending the positive patch to eight positive residues, Cse1(3R), did not appreciably change

the duration of the diffusion traces (8.9 ± 2.2 sec, N=100) and also did not affect the binding affinity for target DNA (**Figures S2**). To further probe the role of Cse1 in promoting Cascade diffusion, we developed a sortase-based transpeptidation strategy to fluorescently label the Cse1 subunit alone, or in complex with Cascade (**Figure S3**). Fluorescent Cse1 could bind and diffuse on DNA, with the longest Cse1 binding events occurring on DNA regions with the highest PAM density (**Figure S3D-F**). Cse1 diffusion trajectories were shorter than those for the Cascade complex at identical ionic strength, suggesting that Cascade also contributes secondary non-specific DNA interactions (**Figure S3E**). A positive groove in the *Tfu*Cse2 subunit is positioned to interact with DNA in the Cascade-crRNA structure and may contribute additional stabilization during target search on non-specific DNA[27,48–50]. Taken together, this data shows that the positive channel formed on the surface of *Tfu*Cse1 is sufficient to promote facilitated diffusion and efficient target recognition by *Tfu*Cascade.

Figure AI.1. Cse1 promotes facilitated diffusion of the Cascade surveillance complex along DNA.

(A) Top: structure of the *T. fusca* (*Tfu*) Cascade surveillance complex (PDB ID: 5U0A). Cascade consists of a crRNA, Cse1(blue/green), two Cse2s (yellow/tan), Cas5 (pink), six Cas7s (light/dark gray), and Cas6e (purple). An epitope on the C-terminus of Cas6e was used for fluorescent labeling (star). Bottom: structure of *Tfu*Cse1 highlighting positive patch residues (purple). Positive patch residues that are evolutionarily conserved, but are neutral in *Tfu*Cse1 are shown in orange. (B) Illustration of double-tethered DNA curtains. A lipid bilayer is deposited on a quartz slide with a microfabricated chrome barrier (B) and pedestals (P). Phage λ DNA is ligated with biotin and digoxigenin (dig)-terminated oligonucleotides and tethered to the lipid bilayer via a biotin-streptavidin linkage. The second DNA end is immobilized on pedestals coated with anti-digoxigenin antibodies. (C) Fluorescent image of double-tethered DNA curtains. DNA (green) is stained with a fluorescent intercalating dye (YOYO-1). Cascade (magenta) binds non-specifically along the DNA substrate. B: barriers; P: pedestals. (D) Illustration (top) and kymographs (bottom) of the indicated Cascade variants scanning DNA for targets via facilitated diffusion. White and red arrows mark DNA binding and release, respectively. (E) Single-particle traces showing six representative Cascade molecules diffusing on DNA. (F) Mean Cascade diffusion coefficients as a function of the ionic strength. N > 45 molecules for all conditions. Error bars: S.E.M. The linear fit (red line) estimates 0.93±0.43 (Avg ± 95% C.I.) Coulombic interactions are disrupted at increasing ionic strength. (G) DNA-binding lifetimes of each Cascade variant. The data was fit to a single exponential decay (solid lines). Half-lives ± 95% C.I. is calculated from the fit. (H) Cascade target binding affinities, as measured via electrophoretic mobility shift assays. Mean and S.D. are calculated from at least three replicates.

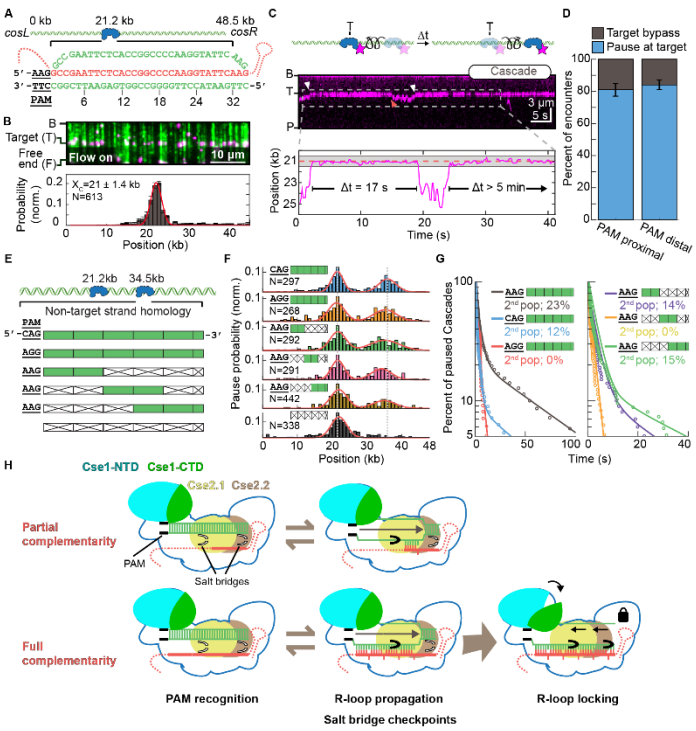


AI.4 Cascade samples potential targets via two transient intermediates

Next, we determined how diffusing Cascade molecules recognize full and partially complementary DNA targets (**Figure AI.2**). Incubating Cascade with the target-containing DNA results in complexes that remain bound at the target site for > 1,900 seconds, indicating full R-loop propagation (**Figures AI.2 & S4**). Direct observation of target recognition dynamics revealed that ~80% of Cascade-target encounters (N=313 encounters) resulted in pausing events longer than 800 ms (**Figure AI.2C**). Diffusing Cascade only pauses at full or partial targets; we did not observe pausing on PAM-rich, but otherwise non-specific DNA (see below). Cascade first recognizes the PAM via the Cse1 subunit, followed by directional extension of the R-loop away from the PAM and along the crRNA[51,52,28,22,42]. However, diffusing Cascade can encounter the target in two polarities—with Cse1 positioned to recognize the PAM and the crRNA oriented in the correct direction for R-loop propagation, or with the crRNA in the opposite orientation relative to the target DNA. Therefore, we also determined whether Cascades that encounter the target from the PAM-proximal or distal sites impact the target recognition frequency. Remarkably, complexes that approach from the PAM-proximal side were just as likely to pause at the target site as those that approach from the PAM-distal end (**Figure AI.2D**). Moreover, after the pauses, Cascade was equally likely to depart from the target site in either PAM-proximal or distal direction (**Figure S4C**). These data are consistent with microscopic hopping during facilitated diffusion. Hopping allows Cascade, and likely other site-specific DNA binding proteins, to sample potential target sites with both polarities, ensuring efficient target recognition.

Figure AI.2. Cascade transiently samples target sequences via PAM-dependent R-loop propagation and seed-distal complementarity.

(A) Illustration of a DNA substrate with a single Cascade target inserted 21.2 kb away from the *cosL* DNA end. The target DNA strand is shown base-paired to the crRNA (red). Numbers indicate flipped out R-loop bases. (B) Top: Image of Cascade (magenta) bound to the target sequence on a single-tethered DNA curtain (green). Bottom: histogram of Cascade binding along the DNA substrate shows a strong preference for the target site. The red line indicates a fit to a Gaussian curve with the center and standard deviation of the fit (error bar) indicated in the figure. (C) Top: illustration and kymograph of a diffusing Cascade molecule transiently pausing at the target site. The white and red arrows indicate the beginning and end of a pause, respectively. Bottom: single-molecule tracking indicates that Cascade pauses twice at the target site (dashed line). The gray band indicates the experimental uncertainty in defining the target site. (D) Cascade pauses with equal frequency at the target regardless of whether it approaches from the PAM-proximal or PAM-distal side (N=27 Cascade molecules; 227 pauses). Error bars are generated from bootstrapping. (E) Schematic of six DNA substrates containing a second Cascade target 34.5 kb away from the *cosL* DNA end. The second targets encode either an altered PAM or segments of the target DNA that are mismatched (white boxes) or complementary (green boxes) with the crRNA. The bottom DNA substrate does not encode any homology to the crRNA and is included as a negative control. (F) Pausing probability of Cascade on the six DNA substrates described in (E). Pausing distributions are fit to two Gaussians (red) and recover both target positions (dotted grey lines). N: number of pauses. (G) Cascade pause durations on the substrates shown in (E). In all but two cases, the data required a bi-exponential fit (solid lines). The magnitude of the second population of the two exponentials is reported. N > 95 pauses for all experiments. (H) Model for target recognition by diffusing Cascade surveillance complexes. The top row represents a target with partial complementarity, and the bottom row a target with full complementarity. Cse1 interacts with the PAM to begin directional unwinding of the DNA duplex. Top row: the extending R-loop is partially stabilized by salt bridges in *Tfu*Cse2.1 and Cse2.2 (black closed gates), but eventually collapses, causing Cascade to leave the target. Bottom: complete R-loop extension locks Cascade onto the DNA target, triggers a conformational change in Cse1, and promotes Cas3 binding (not shown).



Cascade is proposed to engage potential target DNA sites via a series of sequential steps that include PAM recognition, melting of a DNA bubble, propagation of the R-loop past a critical 8-10 nt 'seed' region, and conformational locking[6,22,26–28,51,52]. To further probe this series of steps, we constructed DNA substrates that included a second target site with altered PAMs or partial sequence complementarity to the crRNA (**Figures AI.2E and S4D**). Cascade pausing at these partial target sites required both a PAM as well as a crRNA-complementary segment of target DNA. Surprisingly, scrambling the seed region only resulted in a 50% reduction of paused Cascade molecules relative to the perfect target sequence (**Figure AI.2F**). This suggests that Cascade can transiently recognize PAM-distal target DNA independently of the seed. Next, we observed how long Cascade remained associated with each of the PAM variants and partial target sequences (**Figure AI.2G, left**). Cascade pause times were best described by a bi-exponential fit with a short, t_1=1-3 sec, and a longer, t_2~50 sec, half-life. The PAM controlled the duration and relative amplitude of the shorter timescale (t_1), but not the duration of t_2. The highest DNA-binding affinity (and strongest interference) PAM (5'-AAG) resulted in the longest t_1 pause duration t_1=2.8 ± 0.1 sec (N=656 pauses). In contrast, intermediate interference 5'-CAG and weakest interference 5'-AGG PAMs had short t_1 pauses (t_1=1.5 ± 0.1 sec; N=105 and t_1=2.4 ± 0.4 sec; N=96 pauses, respectively). Moreover, the weakest 5'-AGG PAM pause durations were best described by a single, short exponential decay without a long-lived state (t_2). Next, we determined the pause duration for Cascade on a series of targets that had the strongest PAM (5'-AAG), but contained mismatches between the crRNA and the first, second, and third segments of the target DNA (**Figure AI.2G, right**). All DNA substrates still exhibited a short pause, t_1=~1-2 sec. The second pause duration, t_2, was ~2.6 fold shorter than the perfect target

for substrates with PAM-proximal and distal complementarity, but was virtually non-existent when the complementarity was moved to the middle segment. These data show that complementarity in the PAM-proximal 'seed' region is sufficient to induce a long-lived pause on the partial target as the R-loop directionally propagates away from the PAM. Unexpectedly, PAM-distal complementarity is also sufficient for a long-lived Cascade pause. Taken together with our recent structural work of *Tfu*Cascade R-loop intermediates[53], these results suggest the model summarized in **Figure AI.2H**. The identity of the PAM and the first few PAM-proximal nucleotides initiate a short (1-3 sec) pause. This pause is likely necessary for Cse1 to insert an aromatic wedge into the PAM-proximal DNA duplex and melt a bubble in the target DNA[26]. R-loop propagation is reversible, even on the complementary target DNA. Extension of the R-loop past two Cse2 salt bridges further stabilize the R-loop intermediate[53]. Finally, conformational locking of the entire Cascade complex re-orients the Cse1 N- and C-terminal lobes for Cas3 recruitment and downstream interference and primed acquisition.

AI.5 Translocating Cascade/Cas3 complexes generate tension-sensitive DNA loops

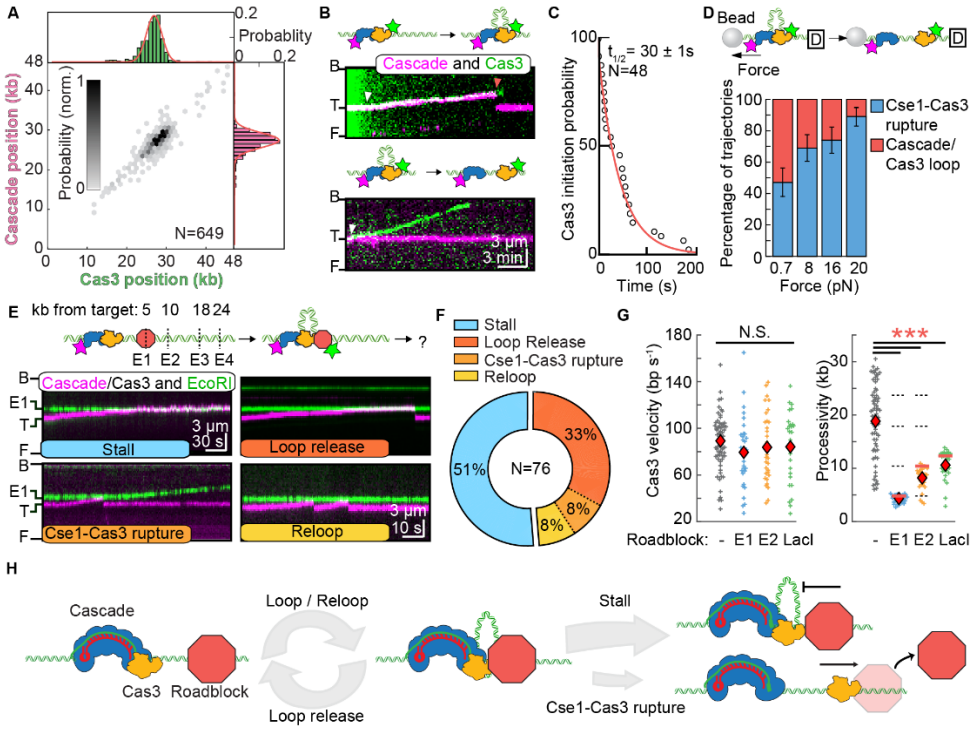
Primed acquisition and interference both require the concerted activities of Cascade and the Cas3 nuclease/helicase. Therefore, we next determined the mechanism of *Tfu*Cas3 recruitment and translocation by imaging Cascade, Cas3, and the ssDNA product. For fluorescent imaging, an ATTO647N dye was directly conjugated to the C-terminus of Cas3 via sortase-mediated transpeptidation (**Figure S5**). Labeling Cas3 with a small C-terminal organic fluorophore was essential because the N-terminus of Cas3 interacts with *Tfu*Cas1-Cas2 (data not shown).

Fluorescent Cas3 localized exclusively to target-bound Cascade and remained stationary on the single-tethered DNA substrates with AMP-PNP, a non-hydrolysable ATP analog (**Figures AI.3A and S5B-D**). These findings are consistent with Cascade loading Cas3 onto the target DNA. In the presence of 1 mM ATP, Cas3 translocated towards the DNA tethering point, as expected for the 3'→5' directionality of the Cas3 helicase domain on the non-target strand (**Figure AI.3B & S5**)[54]. Remarkably, Cascade remained associated with the translocating Cas3 in 47% of all trajectories (**Figure AI.3B**). In the remaining trajectories, Cascade and Cas3 fluorescent signals separated within a single frame (< 200 ms), suggesting a rupture between Cascade and Cas3 that was rapid and stochastic. After rupturing from Cas3, Cascade returned to its initial position at the target DNA site while Cas3 continued to translocate along the DNA substrate (**Figure AI.3B, top**). The co-translocation of the Cascade/Cas3 complex and instantaneous Cascade return to the target site is consistent with a looped DNA intermediate produced during DNA translocation, as has been proposed in other single molecule studies of *E. coli* Type I-E system ([41] and **Loeff *et al***). Tracking individual trajectories revealed an initiation phase where Cas3 showed short bursts of translocation that were largely below our spatial resolution (**Figure S6C** and **Loeff *et al***). Initiation lasted for 30 ± 0.8 s (N=48), followed by processive movement along the DNA substrate (**Figure AI.3C**). Cas3 translocated along the DNA substrate with a mean processivity of 19 ± 7kb (N=68, error bars denote S.D.) at a velocity of 89 ± 25 bp s^{-1} (N=68). Cas3 interacts with the Cse1 subunit of the Cascade complex[21,54,27]. Guided by previous findings that Cas3 interacts with the Cse1 subunit of Cascade and observation that Cse1 and Cas3 are fused in other type I-E systems, we tested whether Cse1 is associated with translocating Cas3 after Cascade release[31,41]. Concurrent dual-color imaging of both Cse1 and Cas6e in a dual-labeled

(ATTO647N)Cse1-Cascade complex revealed that Cse1 always remained associated with Cascade as Cas3 translocated away from the effector complex (**Figure S5E**). These results provide direct evidence for retention of Cse1 in the Cascade effector complex after Cas3 loading and translocation.

Figure AI.3. Processive translocation by the Cascade/Cas3 complex is impeded by DNA-binding proteins.

(A) Histograms of Cas3 (top), Cascade (right), and their joint DNA-binding probability (center) indicate that Cascade loads Cas3 at the target site. (B) Top: illustration and kymograph of a translocating Cascade/Cas3 complex. Cascade remains associated with the target, causing a DNA loop to accumulate during Cas3 translocation. Bottom: Cas3 translocating independently of Cascade. White arrows: initiation of translocation; red arrow: Cascade/Cas3 separation. (C) Cas3 initiates translocation after a 30 ± 1 second pause (N=48). The pause data was fit to a single exponential decay (solid line) to calculate the half-life. Error indicates 95% C.I. (D) Top: experimental configuration for force-dependent Cas3 translocation experiments. The free DNA end was conjugated to a 1 µm paramagnetic bead and hydrodynamic force was applied via buffer flow. Increasing tension on the DNA also increases the frequency of independent Cas3 translocation events, suggesting rupture between the Cse1 and Cas3 protein-protein contacts. (E) Top: illustration of the protein roadblock DNA substrate. The substrate encodes four EcoRI binding sites, E_1 to E_4, positioned 4.8 kb, 10.4 kb, 17.9 and 23.7 kb upstream of the Cascade target. The hydrolytically defective EcoRI(E111Q) was used as a model protein roadblock. Bottom: kymographs showing outcomes of collisions between translocating Cascade/Cas3 complexes (magenta) and EcoRI(E111Q) (green). In all examples, collisions are shown with the first EcoRI(E111Q) bound at E1. (F) Quantification of the collision outcomes observed in (E). (G) Cascade/Cas3 translocation velocities (left) and processivities (right) on naked DNA and with EcoRI(E111Q) or LacI protein roadblocks. Red diamonds indicate the mean of the distribution. For experiments with LacI, the DNA substrate harbored a single ideal LacO site 12.3 kb upstream of the Cascade target. Dashed lines indicate the locations of E1 to E4 and the red lines indicate the location of the first roadblock encountered by Cascade/Cas3. N > 25 for all conditions. The translocation rate was statistically indistinguishable for all conditions (p=0.08, 0.34, 0.42 for EcoRI.E1, EcoRI.E2, and LacI relative to naked DNA, respectively), whereas the processivity was significantly reduced in all roadblock experiments ($p=5.7\times10^{-20}$, 5.9×10^{-19}, 1.6×10^{-12} for EcoRI.E1, EcoRI.E2, and LacI relative to naked DNA, respectively). (H) Model summarizing how Cascade/Cas3 translocates on crowded DNA. Cascade/Cas3 extrude a DNA loop while translocating processively until a collision with a protein roadblock (red octagon). Cascade/Cas3 either slip back or stall at the roadblock. Cas3 can also separate from Cascade and an independently translocating Cas3 can push and evict the roadblock from DNA.



Physical interactions between target-bound Cascade and a moving Cas3 will produce a growing and tension-dependent DNA loop between Cse1 and Cas3 ([41] and Loeff *et al*). To directly visualize these looped DNA intermediates, we used DNA substrates with one fluorescent DNA end positioned either upstream or downstream of translocating Cas3 (**Figure S5F,G**). Consistent with the looping model, Cas3 movement away from the free DNA end also pulls Cascade and the free DNA end at identical rates in the direction of Cas3 translocation (**Figure S5F**, N=10). Alternatively, if the DNA tethering geometry is reversed, then Cas3 translocation will reel in the free DNA end without observable Cascade movement (**Figure S5G**, N=10). Retraction and stochastic release of the free DNA end corresponded with Cas3-dependent translocation and Cse1-Cas3 rupture. In the cell, one or both ends of the foreign DNA are likely to be physically constrained (i.e., to the viral capsid during infection/package or to the transcription/translation machinery during viral replication)[55]. Processive Cascade/Cas3 translocation will thus produce increasing DNA tension as the DNA loop grows. To define the role of DNA tension on Cas3 translocation, we developed a high-throughput assay to measure force-dependent Cascade/Cas3 loop rupture (**Figures AI.3D and S5H**). In this assay, streptavidin is omitted from the lipid bilayer. The chromium pedestals are decorated with anti-DIG antibodies and the DNA is immobilized on the pedestals by its DIG end. The second, biotinylated DNA end is conjugated to 1 μm streptavidin-coated paramagnetic beads. These beads increase the hydrodynamic drag experienced by DNA molecules under mild buffer flow. Increasing the buffer flow rate (hydrodynamic force) correspondingly increases the force applied on the DNA (**Figure S5I**). At an applied force of 0.7 pN, 53% (N=30) of Cascade/Cas3 complexes translocate together for the duration for the entire trajectory. Increasing the applied force resulted in substantially fewer

looped Cascade/Cas3 complexes; only 11% (N=18) of complexes translocated together at 20 pN of applied force (**Figure AI.3D**). We conclude that Cascade/Cas3 interactions rupture as tension accumulates between the moving Cas3 and stationary Cascade.

In the cell, the single-stranded DNA (ssDNA) generated via Cas3 helicase and nuclease activities will be rapidly bound by single-stranded DNA binding protein (SSB). We therefore imaged ssDNA by adding SSB-GFP into the flowcell, and we also determined whether SSB regulates Cas3 activities (**Figure S6**). The intensity of one SSB tetramer on a short ssDNA overhang was used to estimate the number of SSBs associated with each Cas3 (**Figure S6A,B**). Interestingly, SSB-GFP signal accumulated at Cascade/Cas3 complexes prior to processive translocation (**Figure S6C**). These puncta required ATP and were not observed when either Cas3 or ATP were omitted from the flowcells (data not shown). The Cas3 and ATP-dependent generation of ssDNA suggests that Cas3 was translocating distances that were below the ~500 bp resolution of these assays. Consistent with this hypothesis, we occasionally observed repetitive >500 bp Cas3 translocation and slipping that was coincident with a growing SSB-GFP signal (**Figure S6C**). Consistent with our observations, a smFRET-based study with *Ec*Cas3 also observed Cas3 slipping and looped DNA intermediates during translocation (**Loeff *et al***). Moreover, the SSB-GFP signal only increased moderately during processive Cas3 translocation, and never reached full SSB saturation that would be expected if dsDNA were converted to ssDNA, suggesting Cas3 produces short tracts of ssDNA (N=36; **Figure S6D**). These results do not stem from SSB inhibition of Cas3, as neither velocity nor processivity were reduced with SSB added to the flowcell (**Figure S6E**) (**Loeff *et al***). Taken together, our results are consistent with initiation via repetitive rounds of Cas3 slipping and restart, followed by processive Cas3 helicase activity

followed by reannealing of the ssDNA into dsDNA. The Cas3 nuclease domain likely nicks the double-stranded DNA substrate and occasionally produces short tracts of ssDNA that are rapidly coated by SSB.

AI.6 Translocating Cascade/Cas3 is blocked by other DNA-binding proteins

In the cell, DNA is decorated with transcription factors and other DNA-binding proteins. Cas3 will likely encounter these obstacles during processive (>10 kb) translocation. We therefore determined if two site-specific DNA binding proteins—hydrolytically defective EcoRI (E111Q) and Lac repressor (LacI)—influence processive Cas3 translocation (**Figure AI.3E-G**). We first observed Cas3 interactions with fluorescent EcoRI (E111Q) bound specifically to the four EcoRI binding sites on these DNA substrates. The closest two sites, EcoRI.E1 and EcoRI.E2, are 4.8 kb (E_1) and 10.4 kb (E_2) upstream of the Cascade target, respectively (**Figure AI.3E, top**). To assay Cas3 vs. EcoRI(E111Q) collisions, fluorescent Cascade and EcoRI(E111Q) were incubated with the DNA prior to assembling DNA curtains. Cas3 was introduced with ATP, and translocation was monitored via imaging of the Cascade/Cas3 looping complex. EcoRI(E111Q) blocked 100% (N=76/76 molecules) of all Cascade/Cas3 complexes. The most frequent outcome, accounting for 51% of all collisions (N=39/76), was Cascade/Cas3 stalling at the roadblock (**Figures AI.3E,F**). Other outcomes included stalling followed by rupture of the Cas3-Cse1 complex (33%), or loop release and re-looping by the same Cascade/Cas3 complex (8%). Interestingly, after Cas3 leaves Cascade, ~8% (N=6/76) of the Cas3 complexes appeared to push EcoRI(E111Q) off its target site. We never observed roadblock pushing by the entire Cascade/Cas3 complex, suggesting

that Cas3 alone may be more active at removing protein roadblocks. To differentiate the effects of the roadblock from the natural processivity of Cascade/Cas3 on naked DNA, we focused our analysis on Cascade/Cas3 complexes that encountered either of the first two occupied EcoRI(E111Q) binding sites. The observed velocity was statistically indistinguishable to that on naked DNA. However, translocation was blocked by the protein roadblock (**Figure AI.3G, S7**). LacI, located 12.3 kb upstream of the Cascade target, also acted as a strong roadblock to Cascade/Cas3 translocation, with 100% (N=28/28) of collisions resulting in stalling and frequent Cascade/Cas3 loop release (**Figure S7**). In sum, the Cascade/Cas3 complex processively translocates on naked DNA, but is blocked by other DNA-binding proteins (**Figure AI.3H**). Roadblocks may promote Cas3 slipping and re-looping, as has been observed in this study and with *Ec*Cas3 (**Loeff *et al***). Cascade/Cas3 may also stall frequently during translocation on crowded DNA *in vivo*. This stalling may provide additional time for the Cas3 nuclease activity to degrade foreign genetic elements and may also explain the degradation and primed acquisition hotspots reported in prior *in vivo* studies[5,11,38]. Stochastic rupture of the Cse1-Cas3 interface will eventually liberate freely-translocating Cas3 to push and evict roadblocks during CRISPR interference (**Figure AI.3G**).

AI.7 Cas1-Cas2 associates with Cascade/Cas3 in the Primed Acquisition Complex (PAC)

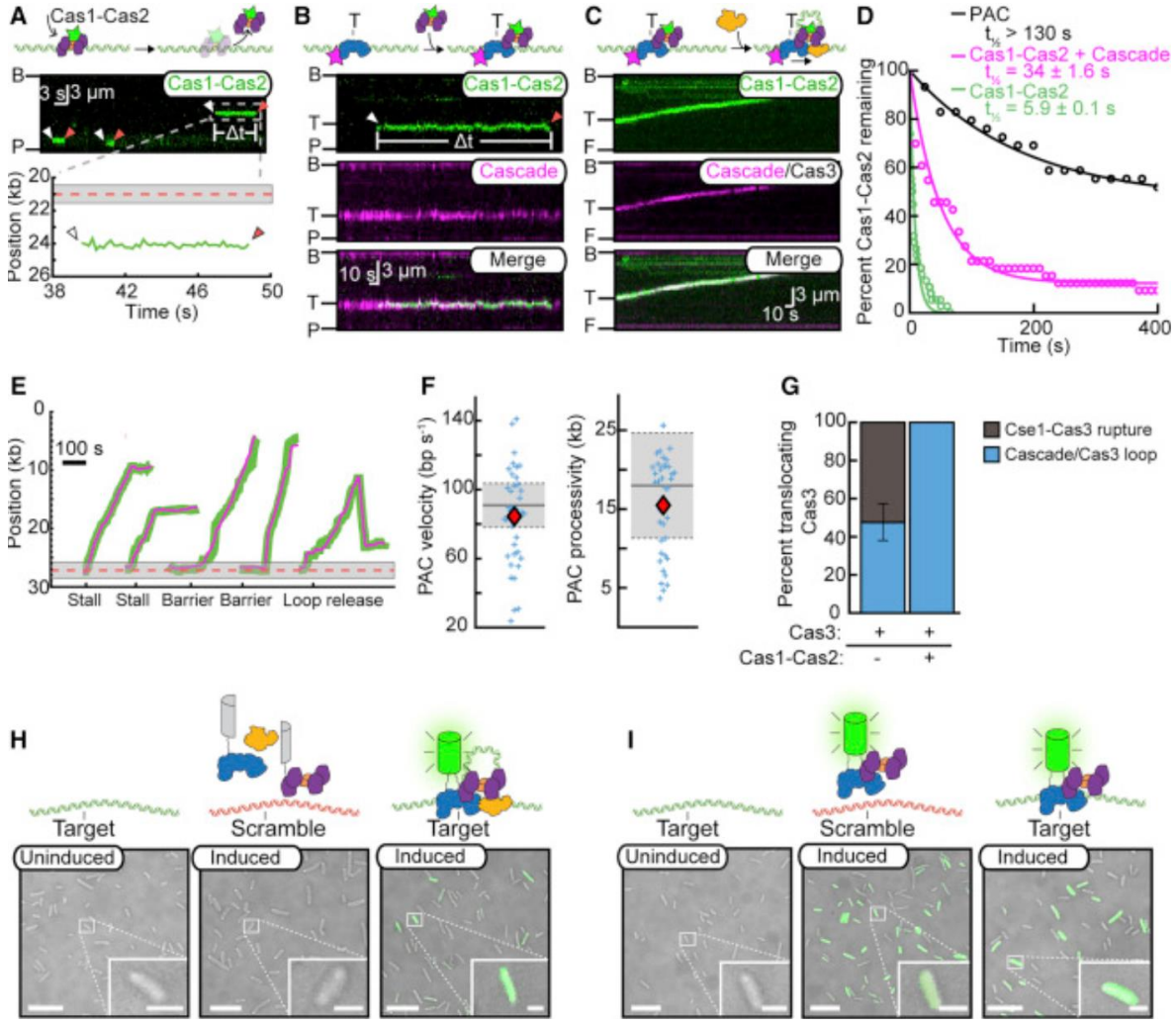
Primed acquisition requires Cascade, Cas3, and the Cas1-Cas2 integrase[5–12]. However, the functions of Cas1-Cas2 in primed acquisition have only been assayed indirectly. Here, we observed the assembly and translocation of a ~710 kDa primed acquisition complex (PAC),

consisting of *Tfu* Cas1-Cas2, Cascade, and Cas3 (**Figure AI.4**). For single molecule imaging, the Cas2 N-terminus was fluorescently labeled via sortase-mediated transpeptidation (**Figure S8**). As expected from biochemical and structural studies of diverse Cas1-Cas2 integrases, *Tfu*Cas1-Cas2 also formed a heterodimer with a $(Cas1)_2$-$(Cas2)_4$ stoichiometry (**Figure S8B**). Cas1-Cas2 transiently bound the DNA substrate with a half-life of ~5.9 ± 0.1 seconds (N=38) (**Figures AI.4A and AI.4D**) and lacked a discernable DNA sequence preference (**Figure S8D**). We next sought to determine how Cas1-Cas2 interacts with the Cascade surveillance complex (**Figure AI.4B**). Three lines of evidence indicated that Cas1-Cas2 forms a long-lived complex with both target-bound and diffusing Cascade complexes. First, Cas1-Cas2 co-localized with Cascade that was pre-loaded on the target site, and the lifetime of this Cas1-Cas2 on DNA increased ~5.8-fold relative to Cas1-Cas2 in the absence of Cascade (**Figures AI.4B and S8E**). Second, pre-incubating fluorescent or unlabeled Cascade with fluorescent Cas1-Cas2, resulted in Cascade/Cas1-Cas2 complexes that diffused on non-specific DNA and could recognize the Cascade target sequence (**Figure S8F**). Third, Cascade could be pulled down with bead-immobilized *Tfu*Cas1-Cas2 (**Figure S8G**). Next, unlabeled Cas3 was added to the pre-assembled Cascade/Cas1-Cas2 sub-complex and the entire primed acquisition complex (PAC) was imaged via dual-color illumination. Directional translocation of the PAC away from the target site confirmed the presence of Cas3 (**Figure AI.4C**). Importantly, the PAC remained stationary when ATP was substituted for the non-hydrolyzable AMP-PNP in the imaging buffer (data not shown). All translocating PACs retained Cas1-Cas2 for the duration of the entire trajectory (N=40), indicating that Cas1-Cas2 is further stabilized within the PAC (**Figures AI.4D & AI.4E**) relative to the Cascade/Cas1-Cas2 sub-complex. All translocating PACs moved towards the DNA tether at a mean velocity of 84 ± 28 bp

s^{-1} (N=40; error indicates S.D.), which was statistically indistinguishable from the velocity observed for Cascade/Cas3 (**Figure AI.4F**). In contrast, the PAC processivity was 25% lower than the Cascade/Cas3 complex (15.5 ± 5.6 kb for the PAC, N=40; p=0.015 relative to Cascade/Cas3). Whereas ~50% of Cascade/Cas3 complexes eventually showed Cse1-Cas3 rupture and independent Cas3 translocation, we did not see any independently translocating Cas1-Cas2/Cas3 sub-complexes under identical force and imaging conditions (**Figure AI.4G**, N=40). Taken together, these results suggest that the Cas1-Cas2 is a core subunit of PAC, where it is stabilized by direct interactions with Cascade. Additional interactions between Cas1-Cas2 and Cas3, as well as the forked DNA that emerges from the Cas3 exit channel also likely contribute to Cas1-Cas2 retention in the PAC.

The formation of the PAC in vivo was also tested via BiFC between Cascade and Cas1 in the presence of Cas3 and target DNA (Figure 4H). Induction of all PAC components produced a fluorescent signal between Cas1-Cas2 and Cascade, but only in the presence of a high-affinity target. In contrast, Cas1-Cas2 bound to Cascade independently of a high-affinity DNA target in the absence of Cas3 (Figure 4I). These data suggest that the PAC organizes around the target DNA and that Cas3 may inhibit the ability of Cas1-Cas2 to bind Cascade in the absence of a target DNA. Taken together, our results demonstrate that Cas1-Cas2 is a core subunit of the PAC, where it is stabilized by direct interactions with Cascade. Additional contacts between Cas1-Cas2 and Cas3, as well as the forked DNA that emerges from the Cas3 exit channel may contribute to Cas1-Cas2 retention in the PAC.

Figure AI.4. Cas1-Cas2 forms a primed acquisition complex (PAC) with Cascade and Cas3.
(A) Illustration (top), kymograph (middle), and quantification (bottom) showing Cas1-Cas2 randomly sampling DNA via 3D collisions. White arrows: Cas1-Cas2 binding, red arrows: Cas1-Cas2 dissociation. Cas1-Cas2 does not show a DNA sequence preference. The dashed red line and gray band represent the Cascade target site, as defined in Figure 2. (B) Illustration (top) and kymographs of Cas1-Cas2 (green) recruitment to Cascade (magenta) at the target sequence. (C) Illustration and a kymograph of the primed acquisition complex (PAC) consisting of Cascade, Cas1-Cas2, and Cas3 processively translocating along the DNA. Cascade (magenta) and Cas1-Cas2 (green) are fluorescently labeled while the presence of dark Cas3 is observed via translocation of the entire complex. (D) DNA-binding lifetimes of Cas1-Cas2 on DNA (blue), as part of the Cascade/Cas1-Cas2 sub-complex (red), and the PAC (green). The data is fit to a single exponential decay. A constant was also included in the Cascade/Cas3 and PAC fits. Error: 95% C.I. (E) Representative traces of the PAC translocating on DNA. Cascade (magenta) and Cas1-Cas2 (green) are fluorescently labeled. The target sequence is shown as a dashed red line and solid gray band. (F) The mean PAC velocity (red diamond) was statistically indistinguishable from Cascade/Cas3 (N≥39 for all datasets; p=0.34). Mean PAC processivity was reduced compared to Cascade/Cas3 (p=0.015). Red diamonds indicate the mean of the PAC distribution. The mean and S.D. of the Cascade/Cas3 distributions are indicated by the solid and dashed lines, respectively. (G) Left: The PAC translocates exclusively via a DNA looping mechanism. Right: termination outcomes for translocating Cascade/Cas3 and PAC complexes. Error bars generated via bootstrapping. (H) BiFC assay showing the PAC forms in vivo. (I) Cascade interacts with Cas1-Cas2 without a target DNA. Scale bars, 10 µm and 2 µm for the insets.



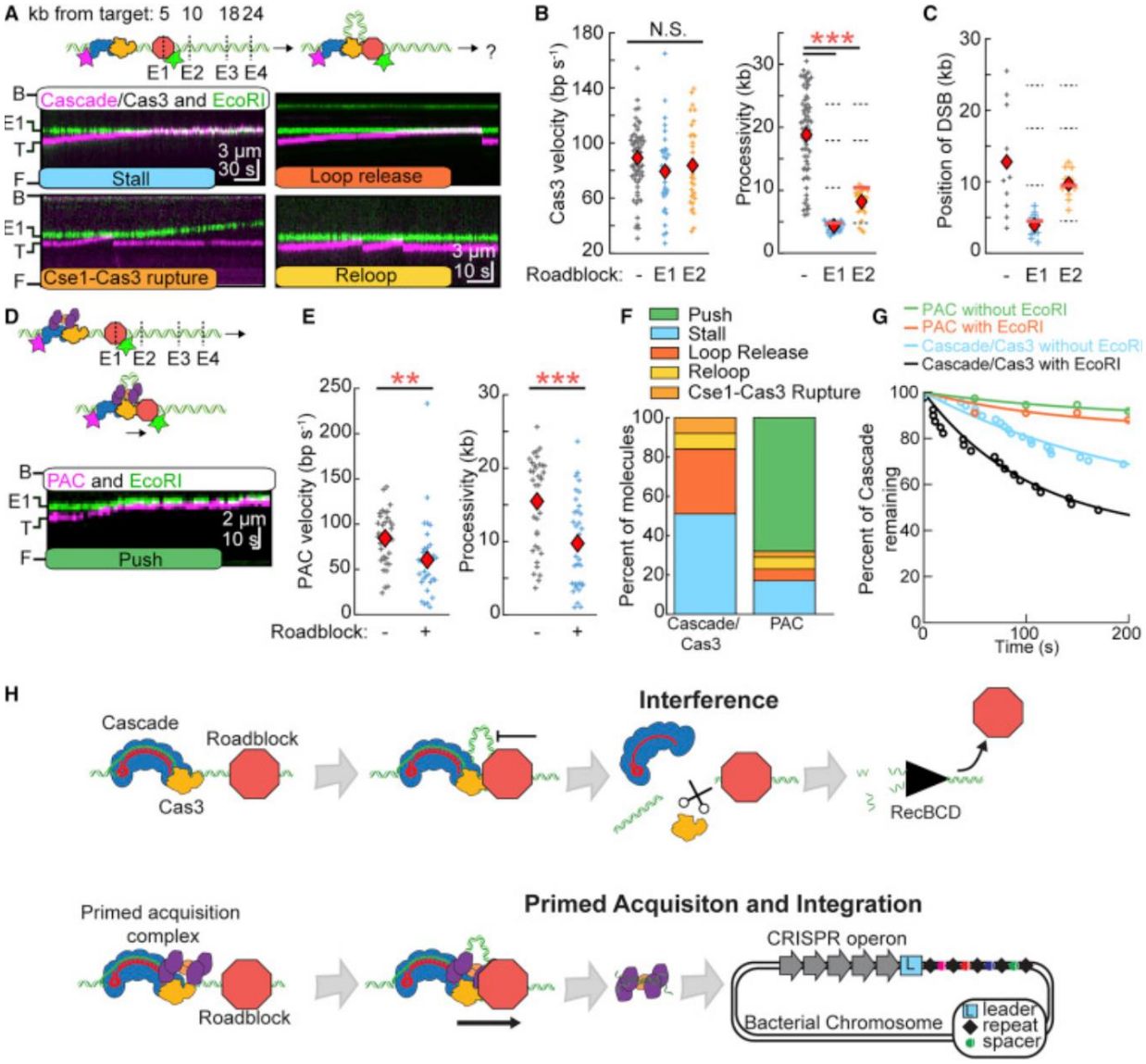
AI.8 Cascade/Cas3 Stalls and Causes DNA Breaks after Colliding with Other DNA-Bound Proteins

Cas3 likely encounters RNA polymerases (RNAPs), transcription factors, and other DNA-binding proteins during processive (>10 kb) translocation. We therefore determined the outcomes of collisions between Cas3 and three site-specific DNA binding proteins—hydrolytically defective EcoRI (E111Q), Lac repressor (LacI), and stalled EcRNAP (Figures 5 and S6). EcoRI (E111Q), LacI, and RNAP bind their target sites with pM-nM affinity and are frequently used as model roadblocks on DNA (Finkelstein and Greene, 2013). We first observed Cas3 interactions with fluorescent EcoRI (E111Q), which bound specifically to four EcoRI binding sites on the DNA **(Figure AI.5A, top)**. To assay Cas3 versus EcoRI (E111Q) collisions, fluorescent Cascade and EcoRI (E111Q) were incubated with the DNA prior to assembling DNA curtains. Cas3 was introduced with ATP, and translocation was monitored via imaging of the Cascade/Cas3 looping complex. EcoRI (E111Q) blocked 100% (n = 76/76) of all Cascade/Cas3 complexes. The most frequent outcome, accounting for 51% of all collisions (n = 39/76), was Cascade/Cas3 stalling at the roadblock (Figures 5A and 5F). Other outcomes included stalling followed by a single-frame release of Cascade/Cas3 back to the initial target site (33%), or re-looping by the same Cascade/Cas3 complex (8%). In the rare event of Cas3 dissociation from Cascade before collision with the roadblock, the freely moving Cas3 could push EcoRI (E111Q) off its target site. We never observed roadblock pushing by the entire Cascade/Cas3 complex, suggesting that Cas3 alone may be able to remove protein roadblocks. To differentiate the effects of the roadblock from the natural processivity of Cascade/Cas3 on naked DNA, we focused our analysis on Cascade/Cas3 complexes that encountered either of the first two occupied EcoRI

(E111Q) binding sites (E1 and E2 in **Figure AI.5**). The observed velocity was statistically indistinguishable from Cas3 on naked DNA. However, translocation was blocked by the protein roadblock **(Figure AI.5B and S6C)**.

Figure AI.5. Differential Outcomes of Translocating Cascade/Cas3 and the PAC at Protein Roadblocks

(A) Top: Illustration of four EcoRI binding sites, E1 to E4, upstream of the Cascade target. Bottom: Outcomes for collisions between translocating Cascade/Cas3 complexes (magenta) and EcoRI(E111Q) bound at E1 (green). **(B)** Cascade/Cas3 translocation velocities (left) and processivities (right) on naked DNA or with EcoRI(E111Q) roadblocks. Red diamonds: mean of the distribution. Dashed lines: locations of E1 to E4. Red line: the location of the first roadblock encountered by Cascade/Cas3. n > 25 for all conditions. Cas3 velocity was statistically indistinguishable for all conditions (p = 0.08, 0.34 for E1 and E2 relative to naked DNA, respectively), whereas the processivity was significantly reduced in all roadblock experiments (p = 5.7 × 10−20, 5.9 × 10−19 for E1 and E2 relative to naked DNA, respectively). **(C)** Position of DSBs induced by Cas3 nuclease activity (n ≥ 10). **(D)** The PAC (magenta) pushes EcoRI(E111Q) (green). **(E)** Velocities (left) and processivities (right) of the PAC in the absence and presence of EcoRI(E111Q). Both velocities and processivities were reduced with a roadblock compared to naked DNA (p = 1.9 × 10−3 and p = 4.9 × 10−5 for velocity and processivity, respectively). **(F)** Outcomes of collisions with EcoRI(E111Q). **(G)** The PAC causes less frequent DSBs on both naked DNA and at a protein roadblock. Error: 95% CI of a single exponential fit. **(H)** Top: Cascade/Cas3 stalls and creates a DSB at roadblocks. Bottom: The PAC can push through roadblocks to acquire additional protospacers.



We also tested two additional protein roadblocks that Cascade/Cas3 would likely encounter in the cell. Lac repressor (LacI) is a bacterial transcription factor that binds its operator site with picomolar affinity and is frequently used as a potent roadblock for DNA motor proteins (Finkelstein and Greene, 2013). LacI, located 12.3 kb upstream of the Cascade target, also blocked Cascade/Cas3 translocation with 100% (n = 28/28) of collisions resulting in stalling and frequent Cascade/Cas3 loop release (Figures S6A–S6C). Finally, we tested conflicts between Cascade/Cas3 and the host RNAP, which is required for early transcription of all foreign DNAs. While Cascade/Cas3 was able to push stalled RNAP (31% of collisions), the most frequent outcome was still Cascade/Cas3 stalling at an EcRNAP (67%) (Figures S6D and S6E). In sum, the Cascade/Cas3 complex processively translocates on naked DNA but is largely blocked by other DNA-binding proteins **(Figure AI.5H)**.

We reasoned that stalled Cas3 may create a double-stranded DNA break (DSB) through concerted nicking via its nuclease activity at the protein roadblock. To test this, we determined the location of Cas3-induced DSBs and the rate of their occurrence with and without the EcoRI roadblock. In the single-molecule assay, DSBs are visualized as a sudden (single-frame) shortening of the DNA molecule along with a loss of the Cascade/Cas3 signal, or by visualization of the cleaved DNA via a DNA intercalating dye (YOYO-1). The lifetime of Cascade/Cas3 on DNA was significantly shorter in the presence of the EcoRI(E111Q) roadblock relative to naked DNA **(Figure AI.5G)**. Cascade dissociation occurred simultaneously with DNA cleavage and required the addition of Cas3 and ATP (Figure S6F). In the absence of any protein roadblocks, DSBs were distributed throughout the DNA. However, Cas3-induced DSBs were predominantly at the

EcoRI.E1 and EcoRI.E2 sites when EcoRI(E111Q) was deposited on the DNA **(Figure AI.5C)**. These results indicate that stalled Cascade/Cas3 complexes cleave DNA at protein roadblocks. The resulting free DNA end may then be further processed by RecBCD and other host nucleases.

AI.9 The PAC Pushes through DNA-Binding Proteins to Search for Downstream Protospacers

Primed acquisition can occur kilobases away from the Cascade target site, indicating that the PAC is also likely to encounter protein obstacles as it translocates on DNA (Semenova *et al*., 2016). Therefore, we tested how the PAC responds to the EcoRI(E111Q) and stalled RNAP protein roadblocks. We first incubated Cascade and EcoRI(E111Q) with the DNA substrate. Next, fluorescent Cas1-Cas2 was injected into the flowcell, followed by Cas3 and 1 mM ATP. Translocation of the PAC was observed as directional movement of Cascade or Cas1-Cas2 away from the target site. The most common outcome of PAC-EcoRI(E111Q) collisions was pushing of the roadblock away from its high-affinity binding site (68% of molecules; n = 24 out of 35) (Figures 5D and 5F). This outcome was markedly different from the Cascade/Cas3-EcoRI(E111Q) collisions, which always blocked translocation **(Figure AI.5F).** Although the PAC could push EcoRI(E111Q), its velocity and processivity decreased significantly relative to the PAC on naked DNA ($p = 1.9 \times 10{-}3$ relative to PAC and $p = 4.9 \times 10{-}5$ relative to PAC, respectively) (Figure 5E). The PAC lifetime was essentially unchanged in the presence of protein roadblocks, suggesting that Cas3-induced DSBs were also significantly downregulated in the context of the PAC (Figure 5G). The PAC could also push promoter-engaged RNAP 63% of the time, suggesting that the PAC is likely able to strip diverse protein roadblocks from cellular DNA (Figure S6). The ability of the PAC to push through protein roadblocks explains the acquisition of additional protospacers relatively far from the Cascade target site (Semenova *et al*., 2016).

AI.10 Discussion

Here, we directly observe the first steps of target recognition and processing by the *Tfu* Type I-E CRISPR-Cas system **(Figure AI.6)**. An evolutionarily conserved positive patch on the outer surface of Cse1 and positive residues in Cas7 promote facilitated diffusion of Cascade during target search. Neutralizing mutations in these positive patches reduce the lifetimes of diffusing Cascade complexes on non-specific DNA and decrease the *in vivo* interference efficiency. Facilitated diffusion is likely a conserved search mechanism among all CRISPR systems (206, 207). Cascade target recognition and stable R-loop locking proceeds via at least two temporally distinct intermediates. The first of these intermediates initiates PAM-proximal opening of the DNA bubble and sampling of the target DNA "seed" region. The second, longer-lived intermediate includes R-loop propagation and additional stabilization via Cse2 salt-bridges. Complexes that cannot fully recognize the R-loop dissociate from the DNA target and continue to scan for targets. After target recognition, Cascade recruits Cas3 helicase/nuclease and the Cascade/Cas3 complex translocates in a 3' to 5' direction on the non-target strand. Cascade remains associated with the target, causing a DNA loop to develop between Cas3 and a target-bound Cascade. This protein interaction ruptures in a stochastic and force-dependent manner, with Cas3 occasionally translocating independently of Cascade. The Cascade/Cas3 complex is highly processive on naked DNA but is blocked by other DNA-binding proteins. Cascade/Cas3 stalling at protein roadblocks allows for iterative nicking by Cas3 and subsequent cleavage of the DNA strand. The resulting DSB can then be further processed by RecBCD and other host nucleases. In contrast, freely moving Cas3 can push protein roadblocks from their DNA-binding

sites. Clearing protein roadblocks by Cas3 could improve the interference efficiency on crowded DNA.

Primed acquisition also requires the Cas1-Cas2 integrase. Here, we provide the first direct evidence that Cas1-Cas2 is stabilized on DNA via physical interactions with Cascade. Cascade forms the keystone of the PAC, as Cas3 and Cas1-Cas2 both require Cascade for stable association with the target DNA. Our data suggest that the PAC can assemble via two routes that include initial recruitment of either Cas3 or Cas1-Cas2 to target-bound Cascade, followed by addition of the remaining sub-complex **(Figure AI.6)**. Further support for this assembly comes from the type I-F system, where Cas3 is expressed as a direct fusion with Cas2.

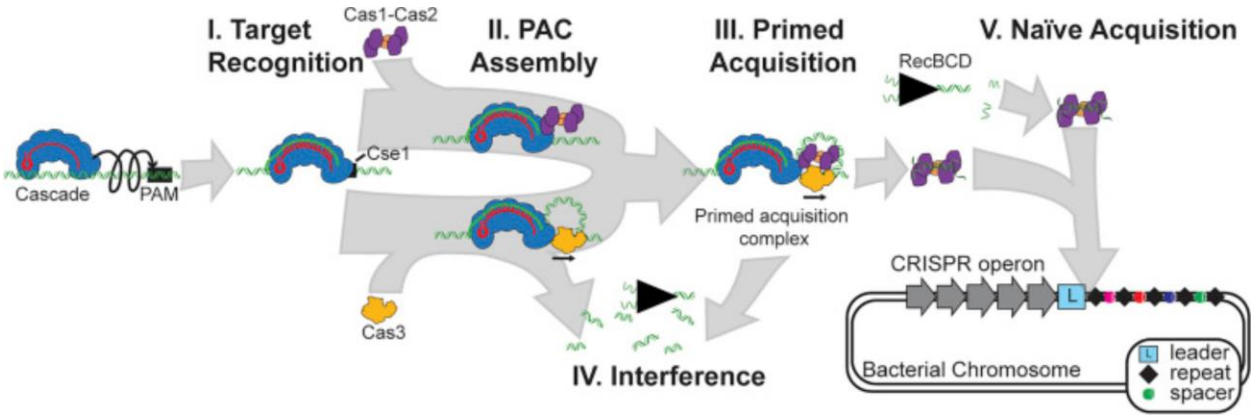
Finally, we demonstrate that the PAC can displace other DNA-binding proteins as it searches for downstream protospacers. Cas1-Cas2 harbors a PAM-decoding center, initially identified in the structure of the EcCas1-Cas2 complex, that is also conserved in TfuCas1-Cas2 (Data S1) (89). The Cas1-Cas2 PAM decoding center may be able to scan, capture, and excise foreign DNAs as they emerge from Cas3 within the PAC. This would likely involve the Cas1 nuclease, as the Cas2 nuclease is structurally occluded and dispensable for integration in vivo (21, 89, 205). Alternatively, further processing by RecBCD and other host nucleases may produce short DNA fragments for integration by Cas1-Cas2 nuclease.

Two models have recently been proposed to account for how interference and primed acquisition are coordinated. One model suggests that Cse1 conformational changes recruit a Cas3/Cas1-Cas2 sub-complex during primed acquisition (82, 208). Cas3/Cas1-Cas2 then moves bi-directionally on the DNA to acquire new spacers. However, spacers are preferentially

selected from the same strand as the original Cascade target site, suggesting the prime acquisition machinery might processively translocate in one direction along the DNA to acquire additional spacers (95). Additionally, a recent single-molecule magnetic tweezers paper found primed acquisition occurs independently of the Cse1 conformational changes (209). An alternative model suggests Cas3 produces DNA cleavage products that Cas1-Cas2 can further process and integrate into the CRISPR locus (85). Our data reconcile these competing models by showing that Cas1-Cas2 forms a complex with Cascade/Cas3, allowing for Cas3 cleavage products to be positioned for direct uptake by Cas1-Cas2. Additional structural and biochemical studies will be required to address how Cas1-Cas2 selects protospacers during PAC translocation and how these protospacers are subsequently integrated into the bacterial genome.

Figure AI.6. Stepwise Assembly of CRISPR-Associated Sub-complexes in Interference and Spacer Acquisition

(I) Cascade surveils foreign DNA via a combination of facilitated 1D diffusion and hopping. (II) Target-bound Cascade can interact with Cas1-Cas2 and Cas3 to assemble the PAC. (III) The PAC samples DNA for possible protospacers during processive translocation. (IV) Alternatively, Cas3 induces a double-stranded DNA break, likely at a protein roadblock. The free DNA ends may be further processed by RecBCD or other host nucleases to generate pre-spacers for adaptive immunity. (V) In naive acquisition, RecBCD degrades foreign DNA into short oligonucleotide-size fragments. Cas1-Cas2 integrates some of these fragments into the CRISPR locus.



AI.11 Methods

Protein Cloning and Purification

Thermobifidua fusca (Tfu) Cascade[54], *Tfu*Cas3[54], *E. coli (Eco)* SSB[59], *Eco*SSB-GFP[60], *Eco* 3xHA-EcoRI(E111Q)[60], *Eco* 3xHA-LacI[60], sortase variants[61–63] and SUMO protease[64] were purified as described previously. For fluorescent labeling, the Cas6 subunit encoded a 3xFLAG epitope tag[53]. *TfuCse1* variants with mutated positive patch residues were cloned by using QuickChange multi-site mutagenesis (Agilent) using oligos MB75, MB76, MB77 & MB78 and MB79 & MB80 for Cse1(5A) and Cse1(3R) respectively (Table 1). Plasmids harboring mutagenized Cse1 (pIF291 for Cse1(5A) or pIF292 for Cse1(3R)) were used to purify Cascade variants following the same protocol as the wild type complex. For fluorescent Cas2 labeling, three glycines were added at the N-terminus using oligos MB069 and MB070 to generate plasmid pIF212 (NEB Q5 mutagenesis kit). Fluorescent Cas3 was prepared by adding LPETG-TwinStrep to the C-terminus with oligos MB073 and MB074 to generate plasmid pIF218.

*Tfu*Cas3 was also purified using a M9 minimal media excluding trace metals. For this purification, Cas3 containing an N-terminal TwinStrep-SUMO-fusion was expressed from a pET-28b expression vector. Starter cultures were prepared by growing 5mL of LB with 50 μg mL^{-1} kanamycin overnight at 37°C. The starter was then transferred to 100 mL M9 containing 50 μg mL^{-1} kanamycin and grown overnight at 37°C. The 1 L expression cultures of M9 containing 50 μg mL^{-1} kanamycin were seeded with 25mL of the overnight M9 starter and were grown at 37°C to an O.D.600 ~ 0.6. Cultures were induced with 1 mM of Isopropyl β-D-1-thiogalactopyranoside (IPTG), 1 μM $CoCl_2$ was added at this time and the cultures were grown

overnight at 22°C. Cells were pelleted and resuspended in 35 mL of Buffer A (30 mM HEPES [pH 7.5], 150 mM NaCl) to be lysed via sonication. After ultracentrifugation, clarified lysate was placed over a 5 mL Strep-Tactin Superflow 50% suspension (IBA Life Sciences, 2-1206-010) gravity column equilibrated in Buffer A. The column was washed with 100 mL of Buffer A and the protein was eluted with 20 mL of Buffer B (30 mM HEPES [pH 7.5], 150 mM NaCl, 5 mM desthiobiotin). After elution, Cas3 was spin concentrated with a (10 kDa) Amicon Ultra-15 Centrifugal Filters (EMD Millipore, UFC903024) and SUMO protease was incubated with the protein overnight. Cas3 was isolated on a HiPrep Sephacryl S-200 HR column (GE, 17116601) pre-equilibrated in Buffer A. Peak fractions were concentrated to 25 µM and frozen with liquid nitrogen.

Tfu Cas1 and Cas2 were cloned into pET expression vectors containing an N-terminal His_6-SUMO-fusion (pIF201 and pIF202 for Cas1 and Cas2, respectively). Cas1 and Cas2 were purified separately following the same protocol: 1 L of LB supplemented with 50 µg mL^{-1} kanamycin was seeded with 20 mL of overnight culture. Cultures were grown at 37°C to an $O.D._{600}$ ~ 0.6 and induced with 0.5 mM Isopropyl β-D-1-thiogalactopyranoside (IPTG). The temperature was reduced to 18°C and growth continued for 18 hours. After expression, cells were pelleted by centrifugation and resuspended in 35 mL of Nickel Buffer A (20 mM HEPES [pH 7.5], 10 mM Imidazole, 500 mM NaCl). Cells were lysed by a pressure homogenizer, and cellular debris pelleted via ultracentrifugation. The clarified lysate was run over two tandem 1 mL His-Trap HP ion affinity columns (GE, 29-0510-21) pre-equilibrated in Nickel Buffer A. The His-Trap column was washed with 40 mL of Nickel Buffer A and Cas1 or Cas2 was eluted with a 20 mL gradient to 100% Nickel Buffer B (20 mM HEPES [pH 7.5], 500 mM imidazole, 500 mM NaCl). SUMO

protease was added to the reaction in 1:50 molar ratio with Cas1 or Cas2 and the mixture was dialyzed against 2 L of Nickel Buffer A overnight. The Cas1-Cas2 complex was assembled by mixing Cas1 and Cas2 at a 4:1 molar ratio and incubating for 1 hour at 4°C. The complex was resolved over a HiPrep Sephacryl S-200 HR column (GE, 17116601) pre-equilibrated in Gel Filtration Buffer (20 mM HEPES [pH 7.5], 500 mM NaCl, 10% glycerol).

Sortase labeling for single-molecule imaging

Peptide synthesis.

Peptides were synthesized using the Liberty Blue Automated Microwave Peptide Synthesizer (CEM Corporation) using manufacturer-suggested protocols. Analytical HPLC characterization of peptides was performed using an Agilent Zorbax column (4.6 x 250 mm; 10 mL min^{-1}, 5-95% MeCN or MeOH (0.1 % trifluoroacetic acid (TFA) or formic acid (FA)) over 60-90 minutes). A Gemini C18 3.5 micron 2.1 x 50 mm was used for online separation; 0.7 mL min^{-1}, 5-95% MeCN (0.1 % formic acid) in 12 min. An Agilent Technologies Accurate-Mass LC/MS (model #6530) was used for high-resolution mass spectra of purified peptides. All solvents were HPLC grade. LPETGG was synthesized using 100 µmole Fmoc-Gly-Wang resin (NovaBiochem by sequential coupling of the N_{α}-Fmoc-amino acid (^{P}3 Biosystems) (0.2 M, 3 ml) in DMF in the presence of DIC (Chem-Impex Inc.) (1M, 1 mL) and ethyl (hydroxyimino)cyanoacetate (1M, 0.5 mL). After final deprotection, the resin was washed three times with 20 mL DMF (Fisher), AcOH, DCM, and MeOH and dried under vacuum. The peptide was cleaved from resin in TFA, water, and triisopropylsilane (TIPS) (95:2.5:2.5) for 3 hours. TFA was removed by flow of nitrogen, and the peptide precipitated with -20°C diethyl ether. Peptide was purified by preparative HPLC

(gradient elution, 5-95% MeOH in H_2O w/ 0.1% FA). Organic solvents were removed by rotary evaporation. Aqueous remnants were frozen at -70°C and lyophilized overnight.

To make Atto647-LPETGG, 5.0 mg of NHS-Atto647N (Atto-Tec) was added to 4.4 mg LPETGG in 1.0 ml of anhydrous DMF. Next, 3.0 µL of N,N-Diisopropylethylamine (Sigma Aldrich) was added. The reaction was placed on a shaker for 3 hours and monitored by LC/MS. Crude mixture was purified directly by preparatory HPLC (5-95% MeOH in H_2O w/ 0.1% FA). Organic solvents were removed by rotary evaporation, aqueous remnants were frozen at -70 °C and lyophilized overnight. Product was isolated as the formic acid salt.

Fmoc-GGGK was synthesized following the procedure described previously, omitting the final Fmoc deprotection step. Peptide was purified by preparative HPLC (gradient elution, 5-95% MeOH in H_2O w/ 0.1% FA). Organic solvents were removed by rotary evaporation, aqueous remnants were frozen at -70 °C and lyophilized overnight. Procedure for making GGGK-Atto647N was performed similarly to Atto647-LPETGG. The reaction was complete after 3 hours. To the crude mixture was added a solution of 20% piperidine in DMF (1.0 ml) and stirred for 20 minutes, deprotecting the N-terminus. Crude mixture was purified directly by preparatory HPLC (5-95% MeOH in H_2O w/ 0.1% formic acid).). Organic solvents were removed by rotary evaporation, aqueous remnants were frozen at -70 °C and lyophilized overnight. Product was isolated as the formic acid salt

Sortase labeling

For fluorescent labeling, Cse1 and Cas2 were purified with an N-terminal GGG residues after the SUMO tag and Cas3 was purified with a C-terminal LPETGG-TwinStrep motif. Sortase labeling

was optimized for each protein by varying the temperature, labeling time, and sortase variant[61–63]. Cse1 was labeled by incubating 48 μM Cse1 with 50 μM sortase(5M), 10 mM $CaCl_2$, 250 μM (Atto647N)-LPETGG fluorescent peptide, and 60 μM SUMO protease for 12 hours at 4˚C. Immediately following fluorescent labeling, Cse1 was separated from the free peptide and sortase on a Sephacryl S-200 HR column (GE) using glycerol free Gel Filtration Buffer. Fluorescent Cse1 was then reconstituted with the rest of the Cascade complex in a 1:1 ratio through a step-down NaCl dialysis (500 mM NaCl to 150 mM NaCl), and the full complex was isolated using a HiPrep Sephacryl S-200 HR column (GE) with TS Buffer (10 mM Tris-HCl [pH 7.5], 150 mM NaCl, 5 mM DTT).

Fluorescent Cas2 was prepared by cloning a C-terminal GGG purified similar to wild type Cas2 with the following modification. After SUMO proteolysis, 20 μM of GGG-Cas2 was incubated with 100 μM sortase(7M) and 100 μM (Atto647N)-LPETGG fluorescent peptide at 4 °C for 1 hour along with of 5 mM $CaCl_2$. Fluorescent Cas2 was separated from the free peptide and sortase on a HiPrep Sephacryl S-200 HR column with Gel Filtration Buffer. Fluorescently labeled Cas3 was generated by incubating 20 μM of Cas3-LPETGG-TwinStrep with 100 μM sortase(7M) and 100 μM GGGK-(Atto647N) fluorescent peptide at 15 °C for 1 hour along with 5 mM of $CaCl_2$. Fluorescent Cas3 was separated from the free peptide and sortase using a HiPrep Sephacryl S-200 HR column with Gel Filtration Buffer containing 150 mM NaCl.

Antibodies

Cascade was fluorescently labeled with mouse anti-FLAG M2 (Sigma, F3165) via a 3xFLAG epitope tag on the Cas6 subunit[53]. For single-molecule imaging, antibodies were conjugated to

605 nm or 705 nm quantum dots (QDs) following published protocols (Thermo Fisher Scientific) [45,59]. QD-conjugated antibodies were stored in PBS Buffer (pH 7.2, with 2 mM sodium azide) at 4°C. The following antibodies were used for Westerns and co-IP experiments with Cas2, Cas3-6xHis, and Cascade-1xFLAG, respectively: 6xHis Monoclonal Antibody (Albumin Free, Clontech, 631212) and DYKDDDDK Tag Antibody (Cell Signaling Technology, 2368S)

Electrophoretic mobility shift assay (EMSA)

All EMSAs were performed with Cy5-labeled DNA substrates that were generated via PCR with primers CJ1 and CJ2, as described previously[53]. Cascade EMSAs were performed by incubating 0.3 nM of the PCR product with increasing Cascade concentrations (0.13, 0.22, 0.37, 0.62, 1.0, 1.7, 2.9, 4.8, 8 nM for WT and Cascade(3R); 1.8, 4.6, 12, 29, 72, 180, 450 nM for Cascade(5A)) for 30 minutes at 62°C in Binding Buffer (20 mM HEPES [pH 7.5], 150 mM NaCl, 2 mM $MgCl_2$, 1 mM DTT, 0.2 mg ml-1 BSA, 0.01% Tween-20). The reactions were resolved on a 5% native PAGE gel with 0.5X TBE Buffer (45 mM Tris-HCl [pH 8.0], 45 mM boric acid, 1 mM EDTA). Gels were visualized using a Typhoon scanner (GE) and quantified in ImageQuant TL v8.1 (GE). The fraction of bound DNA was fit to the hyperbolic curve to obtain K_d values. All experiments were repeated in triplicate.

DNA substrates for single-molecule microscopy

DNA substrates with mutated target sequences were generated by cloning the mutated targets into helper plasmids pIF152 and pIF153 that had ~200 bp of flanking homology with λ-phage DNA[65]. PCR products containing the large homology arms were recombineered into *E. coli* lysogens and the recombinant DNA purified from packaged phage particles[65]. To functionalize

the DNA ends for single-molecule experiments, we combine 125 µg of purified λ-phage DNA with 2 µM of biotinylated oligos (IF001or IF003 or (Table 1)). For double-tethered DNA curtains, a second dig-labeled oligo was annealed to the second DNA end (oligos IF002 or IF004). After ligation, the reaction was separated over a Sephacryl S-1000 column (GE, #45-000-084) to purify full length labeled DNA. The DNA was stored at 4°C.

Single-molecule fluorescence microscopy and data analysis

All single-molecule imaging was performed using a Nikon Ti-E microscope in a prism-TIRF configuration equipped with a motorized stage (Prior ProScan II H117) containing microfluidic flowells housed in a custom stage adapter. The flowcell was illuminated with 488 nm (Coherent), 532 nm (Ultralasers), and 633 nm (Ultralasers) lasers through a quartz prism (Tower Optical Co.)[40]. A 60x air objective and a custom built microscope stage heater were used to maintain the flowcell near the optimal *Tfu*Cascade temperature.

To prepare double-tethered DNA curtains for single-molecule imaging, 40 µL of liposome stock solution (97.7% DOPC, 2.0% DOPE-mPEG2k, and 0.3% DOPE-biotin; Avanti #850375P, #880130P, #870273P, respectively) was diluted into 960 µL Lipids Buffer (10 mM Tris-HCl [pH 7.8], 100 mM NaCl) and incubated in the flowcell for 30 minutes. Next, 50 ng μL^{-1} of goat anti-rabbit polyclonal antibody (ICL Labs, #GGHL-15A) diluted in Lipids Buffer was injected into the flowcell and incubated for 10 minutes. The flowcell was washed in Imaging Buffer (40 mM Tris–HCl [pH 7.8], 2 mM $MgCl_2$, 0.2 mg mL^{-1} BSA) followed by 10 minute incubation with 5 ng μL^{-1} of digoxigenin monoclonal antibody (Life Technologies, #700772) diluted in Imaging Buffer. Next, 0.1 mg mL^{-1} Streptavidin diluted in Imaging Buffer was injected into the flowcell and incubated

for 10 minutes. Lastly, 12.5 ng μL^{-1} of the biotin- and dig-labeled DNA substrate was injected into the flowcell. Single-tethered curtains were prepared by omitting the anti-rabbit antibody and digoxigenin antibody steps. In all experiments, Imaging Buffer was supplemented with 50 mM NaCl. In experiments using Sortase labeled Cas3 and/or Cas1-Cas2, 10 mL Imaging Buffer was supplemented with 1 mM Trolox (Sigma-Aldrich, #238813-5G), 500 units of catalase (Sigma-Aldrich), 70 units of glucose oxidase (Sigma-Aldrich), and 1% glucose (w/v).

To observe Cascade diffusion and target search, 150 µL of 0.1 nM QD-labeled Cascade in Imaging Buffer supplemented with 50-150 mM NaCl was injected into a flowcell with pre-assembled double-tethered DNA curtains. Excess Cascade was removed and DNA-bound Cascade complexes were imaged at a 200 ms framerate for 10 minutes. In experiments where Cascade was pre-bound to the DNA target, 10 nM Cascade was incubated with 1.3 µg of biotinylated and digoxigenin labeled DNA at 55 °C for 10 minutes, followed by a 10-minute incubation at room temperature. The DNA bound Cascade was then diluted to 1 mL in Imaging Buffer with 50 mM NaCl and injected into flowcells prepared for single or double-tethered DNA curtains. Cascade was then labeled *in situ* by injecting 150 µL of 10 nM anti-FLAG antibody conjugated QDs.

Cas3 translocation was observed by injecting 10 nM Cas3 diluted in 150 µL Imaging Buffer onto single-tethered DNA curtains with Cascade pre-bound to its target DNA. Experiments using ATTO647N-Cas3 used a five second frame rate and a computer-controlled digital shutter (Vincent Associates) on the 633 nm laser to limit Cas3 photobleaching. In these experiments, Cascade was visualized using a spectrally distinct 605 nm QD. Wild type (unlabeled) Cas3 was used in experiments with fluorescent Cse1 or fluorescent Cas1-Cas2 complex.

To observe Cas3-roadblock collisions, Cascade was pre-bound to the DNA target at 55°C. Then 5 nM of 3xHA-EcoRI(E111Q) or 2.5 nM 3xHA-LacI were incubated with the Cascade-DNA substrate on ice for 5 minutes, followed by dilution into 1 mL of Imaging Buffer. The protein bound DNA was then injected into flowcells prepared for single-tethered DNA curtains. All proteins were labeled in situ. HA labeled proteins were labeled by injecting 150 µL of 1 nM of anti-rabbit conjugated Qdots (Thermo Q-11461MP) pre-bound to 0.2 nM anti-HA antibody (ICL Labs, RHGT-45A-Z) diluted in Imaging Buffer. For experiments involving RNAP complexes, Cascade was pre-bound to the DNA at 55°C. E. coli RNAP holoenzyme was fluorescently labeled with a streptavidin-coated QD (Finkelstein *et al*., 2010) and injected into the flowcell in the presence of 25 µM of GTP, 1 mM ATP, and 25 µM UTP. The ATP concentration was higher to support Cas3 translocation. RNAP that was not engaged to the promoter was removed from the DNA by a 700 µL heparin wash (0.2 mg mL-1). Cascade was fluorescently labeled by a QD in situ. Then, 10 nM unlabeled Cas3 was injected and collisions between the RNAP and Cascade/Cas3 complexes were visualized by recording ~10-min movies at 5 frames per second.

For force-dependent experiments, 12 µg of biotinylated and digoxigenin labeled λ-DNA molecules were conjugated to 4 mg of 1 µm superparamagnetic beads (NEB, #S1420S) in Lipids Buffer overnight at room temperature. DNA-conjugated beads were washed 3 times and resuspended in 75 µL of Lipids Buffer. Cascade (30 nM) was pre-bound to 15 µL of DNA conjugated beads at 55°C and cooled to room temperature. DNA was captured in flowcells assembled with liposomes lacking biotinylated lipids and streptavidin. Cascade-bound DNA was injected into the flowcell, and concentrated at the surface with a rare earth magnet for 10 minutes. The DNA bound to digoxigenin antibodies at the chromium barriers. Excess DNA and

beads were flushed out of the flowcell. To initiate Cas3 translocation, 10 nM of Cas3 was injected into this flowcell at 50 µL min^{-1}. The flow rate was subsequently increased to the desired applied force. To calculate the force-dependent elongation of DNA conjugated beads, single particle tracking was used to measure the mean extension of bead-tethered DNA molecules from the chromium barriers at flow rates ranging from 100 to 1200 µL min^{-1}.

To image fluorescent Cas1-Cas2, Cascade was pre-bound to the target and labeled via the 3xFLAG epitope on Cas6, as described above. Fluorescent Cas1-Cas2 was diluted to a final concentration of 1 nM in Imaging Buffer containing 150 mM NaCl, and injected onto single-tethered DNA curtains. Free Cas1-Cas2 was washed out of the flowcell, followed by injection of 10 nM Cas3 when indicated.

Co-immunoprecipitation and Western Blotting

Purified Cas1-Cas2 (225 nM) was incubated with purified Cascade (225 nM) on ice for 30 minutes in Western Buffer (40mM Tris-HCl [pH 8.0], 0.2 mg/mL BSA, 150 mM NaCl, 10% glycerol, 2 mM $MgCl_2$, and 2 units/mL DNase I. To pull-down by TwinStrep-Cas2, the sample was applied to Strep-tactin Superflow 50% suspension beads (catalog# 2-1206-002, IBA). Anti-FLAG M2 Magnetic Beads (catalog# M8823-1ML, Sigma) were used to carry out the reciprocal experiment by pulling-down via Cascade-1xFLAG. The beads were then washed three times with Western Buffer, and the samples removed by adding 3x-FLAG peptide or boiling the beads. Supernatant was resolved on a 15% SDS-PAGE gels and probed by standard Western blotting.

Data Analysis

Fluorescent particles were tracking using an in-house ImageJ script (available upon request). Trajectories were used to calculate the mean-squared displacement and the diffusion coefficients for Cascade, or the velocity and processivity for the Cas3-containing complexes, as described previously [66,59]. Binding lifetimes were fit to either a single exponential decay or a biexpoential decay using a custom MATLAB script (Mathworks R2015b). The biexpoential fits were tested to be appropriate using an *f-test* applied to the survival curve data[59]. For pause analysis, a molecule was considered paused if it stayed within a stationary window for four continuous frames (0.8 secounds). This window was defined as 3-fold the standard deviation (S.D.) of the fluctuations of a stationary Cascade at its target[45]. Pause location was recorded in relation to the pedestal located at the digoxigenin labeled end of the DNA.

Translocating Cas3 was defined as Cas3 that left the target window for at least four continuous frames (> 800 ms). Looping Cas3-Cascade molecules were defined by scoring whether Cascade also left the target window with Cas3. In contrast, independently moving Cas3s were defined by scoring traces where Cascade remained stationary while Cas3 moved away from the target window.

Roadblock collision analysis. Collisons were defined when Cascade fluorescence co-localized with the roadblock (EcoRI or LacI). The roadblock was considered pushed if it moved away from its binding site for four adjacent frames (0.8 seconds).

Cse1 homology modeling. Multi-sequence alignment was performed with the ConSurf evolutionary conservation tool using the HMMR homolog search algorithm, and MAFFT multiple sequence alignment methods[47]. Conservation of positive residues was calculated as

the percentage of a total of 150 divergent Cse1 homolog sequences that had an Arginine, Lysine, and Histidine for each residue aligned against *Tfu*Cse1.

AI.12 Acknowledgments: We are very grateful to Jim Rybarski, Andrew Leal, and Brianna Gonzalez for providing materials. This work was supported by the Welch Foundation (F-1808 to I.J.F.), the US National Institutes of Health grant R01GM124141 to I.J.F. and R35GM118174 to A.K. I.J.F. is a CPRIT Scholar in Cancer Research. L.M. was supported by NIH fellowship F99CA212452.

AI.13 Author Contributions: M.B., K.D., L.M., Y.X., A.K., and I.J.F. conceived the study. M.B., K.D., L.M., Y.X., A.K., and I.J.F. designed the experiments and analyzed the data. M.W.B., K.E.D., Y.X., A.D., and L.R.M. generated key materials and executed the experiments. E.H. and S.D. provided fluorescent peptides. Y.K. provided particle tracking and roadblock bypass software. M.B., K.D., and I.J.F. wrote the paper with input from all other authors.

Author Information: Reprints and permissions information is available online. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to M.B. (maxwellbrown@utexas.edu), K.D. (kaylee.dillard@utexas.edu), or I.J.F. (ifinkelstein@cm.utexas.edu).

AI.14 APPENDIX I REFERENCES

1. Sorek, R., Lawrence, C. M. & Wiedenheft, B. CRISPR-Mediated Adaptive Immune Systems in Bacteria and Archaea. *Annu. Rev. Biochem.* **82,** 237–266 (2013).
2. Barrangou, R. *et al.* CRISPR provides acquired resistance against viruses in prokaryotes. *Science* **315,** 1709–1712 (2007).
3. Mohanraju, P. *et al.* Diverse evolutionary roots and mechanistic variations of the CRISPR-Cas systems. *Science* **353,** aad5147 (2016).
4. Makarova, K. S. *et al.* An updated evolutionary classification of CRISPR-Cas systems. *Nat. Rev. Microbiol.* **13,** 722–736 (2015).
5. Datsenko, K. A. *et al.* Molecular memory of prior infections activates the CRISPR/Cas adaptive bacterial immunity system. *Nat. Commun.* **3,** 945 (2012).
6. Fineran, P. C. *et al.* Degenerate target sites mediate rapid primed CRISPR adaptation. *Proc. Natl. Acad. Sci.* **111,** E1629–E1638 (2014).
7. Li, M., Wang, R., Zhao, D. & Xiang, H. Adaptation of the Haloarcula hispanica CRISPR-Cas system to a purified virus strictly requires a priming process. *Nucleic Acids Res.* **42,** 2483–2492 (2014).
8. Richter, C. *et al.* Priming in the Type I-F CRISPR-Cas system triggers strand-independent spacer acquisition, bi-directionally from the primed protospacer. *Nucleic Acids Res.* **42,** 8516–8526 (2014).
9. Xue, C. *et al.* CRISPR interference and priming varies with individual spacer sequences. *Nucleic Acids Res.* **43,** 10831–10847 (2015).
10. Fagerlund, R. D. *et al.* Spacer capture and integration by a type I-F Cas1–Cas2-3 CRISPR adaptation complex. *Proc. Natl. Acad. Sci.* **114,** E5122–E5128 (2017).
11. Semenova, E. *et al.* Highly efficient primed spacer acquisition from targets destroyed by the Escherichia coli type I-E CRISPR-Cas interfering complex. *Proc. Natl. Acad. Sci.* **113,** 7626–7631 (2016).
12. Jackson, S. A. *et al.* CRISPR-Cas: Adapting to change. *Science* **356,** eaal5056 (2017).
13. Garneau, J. E. *et al.* The CRISPR/Cas bacterial immune system cleaves bacteriophage and plasmid DNA. *Nature* **468,** 67–71 (2010).
14. Terns, M. P. & Terns, R. M. CRISPR-Based Adaptive Immune Systems. *Curr. Opin. Microbiol.* **14,** 321–327 (2011).
15. Barrangou, R. & Marraffini, L. A. CRISPR-Cas Systems: Prokaryotes Upgrade to Adaptive Immunity. *Mol. Cell* **54,** 234–244 (2014).
16. Marraffini, L. A. CRISPR-Cas immunity in prokaryotes. *Nature* **526,** 55–61 (2015).
17. Brouns, S. J. J. *et al.* Small CRISPR RNAs guide antiviral defense in prokaryotes. *Science* **321,** 960–964 (2008).
18. Hochstrasser, M. L. & Doudna, J. A. Cutting it close: CRISPR-associated endoribonuclease structure and function. *Trends Biochem. Sci.* **40,** 58–66 (2015).
19. Carte, J., Pfister, N. T., Compton, M. M., Terns, R. M. & Terns, M. P. Binding and cleavage of CRISPR RNA by Cas6. *RNA N. Y. N* **16,** 2181–2188 (2010).
20. Charpentier, E., Richter, H., van der Oost, J. & White, M. F. Biogenesis pathways of RNA guides in archaeal and bacterial CRISPR-Cas adaptive immunity. *FEMS Microbiol. Rev.* **39,** 428–441 (2015).
21. Hochstrasser, M. L. *et al.* CasA mediates Cas3-catalyzed target degradation during CRISPR RNA-guided interference. *Proc. Natl. Acad. Sci. U. S. A.* **111,** 6618–6623 (2014).
22. Blosser, T. R. *et al.* Two distinct DNA binding modes guide dual roles of a CRISPR-Cas protein complex. *Mol. Cell* **58,** 60–70 (2015).
23. Wiedenheft, B. *et al.* RNA-guided complex from a bacterial immune system enhances target recognition through seed sequence interactions. *Proc. Natl. Acad. Sci. U. S. A.* **108,** 10092–10097 (2011).

24. Jore, M. M. *et al.* Structural basis for CRISPR RNA-guided DNA recognition by Cascade. *Nat. Struct. Mol. Biol.* **18,** 529–536 (2011).
25. Xue, C., Whitis, N. R. & Sashital, D. G. Conformational Control of Cascade Interference and Priming Activities in CRISPR Immunity. *Mol. Cell* **64,** 826–834 (2016).
26. Hayes, R. P. *et al.* Structural basis for promiscuous PAM recognition in type I-E Cascade from E. coli. *Nature* **530,** 499–503 (2016).
27. Xiao, Y. *et al.* Structure Basis for Directional R-loop Formation and Substrate Handover Mechanisms in Type I CRISPR-Cas System. *Cell* **170,** 48–60.e11 (2017).
28. Rutkauskas, M. *et al.* Directional R-Loop Formation by the CRISPR-Cas Surveillance Complex Cascade Provides Efficient Off-Target Site Rejection. *Cell Rep.* (2015). doi:10.1016/j.celrep.2015.01.067
29. Sashital, D. G., Wiedenheft, B. & Doudna, J. A. Mechanism of foreign DNA selection in a bacterial adaptive immune system. *Mol. Cell* **46,** 606–615 (2012).
30. Sinkunas, T. *et al.* Cas3 is a single-stranded DNA nuclease and ATP-dependent helicase in the CRISPR/Cas immune system. *EMBO J.* **30,** 1335–1342 (2011).
31. Westra, E. R. *et al.* CRISPR Immunity Relies on the Consecutive Binding and Degradation of Negatively Supercoiled Invader DNA by Cascade and Cas3. *Mol. Cell* **46,** 595–605 (2012).
32. Huo, Y. *et al.* Structures of CRISPR Cas3 offer mechanistic insights into Cascade-activated DNA unwinding and degradation. *Nat. Struct. Mol. Biol.* **21,** 771–777 (2014).
33. Gong, B. *et al.* Molecular insights into DNA interference by CRISPR-associated nuclease-helicase Cas3. *Proc. Natl. Acad. Sci. U. S. A.* **111,** 16359–16364 (2014).
34. Krupovic, M., Makarova, K. S., Forterre, P., Prangishvili, D. & Koonin, E. V. Casposons: a new superfamily of self-synthesizing DNA transposons at the origin of prokaryotic CRISPR-Cas immunity. *BMC Biol.* **12,** 36 (2014).
35. Wang, J. *et al.* Structural and Mechanistic Basis of PAM-Dependent Spacer Acquisition in CRISPR-Cas Systems. *Cell* **163,** 840–853 (2015).
36. Arslan, Z., Hermanns, V., Wurm, R., Wagner, R. & Pul, Ü. Detection and characterization of spacer integration intermediates in type I-E CRISPR–Cas system. *Nucleic Acids Res.* **42,** 7884–7893 (2014).
37. Levy, A. *et al.* CRISPR adaptation biases explain preference for acquisition of foreign DNA. *Nature* **520,** 505–510 (2015).
38. Staals, R. H. J. *et al.* Interference-driven spacer acquisition is dominant over naive and primed adaptation in a native CRISPR–Cas system. *Nat. Commun.* **7,** ncomms12853 (2016).
39. Gallardo, I. F. *et al.* High-Throughput Universal DNA Curtain Arrays for Single-Molecule Fluorescence Imaging. *Langmuir* **31,** 10310–10317 (2015).
40. Soniat, M. M. *et al.* Next-Generation DNA Curtains for Single-Molecule Studies of Homologous Recombination. in *Methods in Enzymology* (ed. Eichman, B. F.) **592,** 259–281 (Academic Press, 2017).
41. Redding, S. *et al.* Surveillance and Processing of Foreign DNA by the Escherichia coli CRISPR-Cas System. *Cell* **163,** 854–865 (2015).
42. Sternberg, S. H., Redding, S., Jinek, M., Greene, E. C. & Doudna, J. A. DNA interrogation by the CRISPR RNA-guided endonuclease Cas9. *Nature* **507,** 62–67 (2014).
43. Blainey, P. C. *et al.* Nonspecifically bound proteins spin while diffusing along DNA. *Nat. Struct. Mol. Biol.* **16,** 1224–1229 (2009).
44. Wang, F. *et al.* The promoter-search mechanism of Escherichia coli RNA polymerase is dominated by three-dimensional diffusion. *Nat. Struct. Mol. Biol.* **20,** 174–181 (2013).
45. Brown, M. W. *et al.* Dynamic DNA binding licenses a repair factor to bypass roadblocks in search of DNA lesions. *Nat. Commun.* **7,** 10607 (2016).
46. Tay, M., Liu, S. & Yuan, Y. A. Crystal structure of Thermobifida fuscaCse1 reveals target DNA binding site. *Protein Sci.* **24,** 236–245 (2015).

47. Ashkenazy, H. *et al.* ConSurf 2016: an improved methodology to estimate and visualize evolutionary conservation in macromolecules. *Nucleic Acids Res.* **44,** W344–W350 (2016).
48. van Erp, P. B. G. *et al.* Mechanism of CRISPR-RNA guided recognition of DNA targets in Escherichia coli. *Nucleic Acids Res.* **43,** 8381–8391 (2015).
49. Jackson, R. N. *et al.* Crystal structure of the CRISPR RNA–guided surveillance complex from Escherichia coli. *Science* **345,** 1473–1479 (2014).
50. Mulepati, S., Héroux, A. & Bailey, S. Crystal structure of a CRISPR RNA–guided surveillance complex bound to a ssDNA target. *Science* **345,** 1479–1484 (2014).
51. Semenova, E. *et al.* Interference by clustered regularly interspaced short palindromic repeat (CRISPR) RNA is governed by a seed sequence. *Proc. Natl. Acad. Sci.* **108,** 10098–10103 (2011).
52. Szczelkun, M. D. *et al.* Direct observation of R-loop formation by single RNA-guided Cas9 and Cascade effector complexes. *Proc. Natl. Acad. Sci. U. S. A.* **111,** 9798–9803 (2014).
53. Jung, C. *et al.* Massively Parallel Biophysical Analysis of CRISPR-Cas Complexes on Next Generation Sequencing Chips. *Cell* **170,** 35–47.e13 (2017).
54. Huo, Y. *et al.* Structures of CRISPR Cas3 offer mechanistic insights into Cascade-activated DNA unwinding and degradation. *Nat. Struct. Mol. Biol.* **21,** 771–777 (2014).
55. Modell, J. W., Jiang, W. & Marraffini, L. A. CRISPR–Cas systems exploit viral DNA injection to establish and maintain adaptive immunity. *Nature* **544,** 101–104 (2017).
56. Rollins, M. F. *et al.* Cas1 and the Csy complex are opposing regulators of Cas2/3 nuclease activity. *Proc. Natl. Acad. Sci.* **114,** E5113–E5121 (2017).
57. Nuñez, J. K. *et al.* Cas1–Cas2 complex formation mediates spacer acquisition during CRISPR–Cas adaptive immunity. *Nat. Struct. Mol. Biol.* **21,** 528–534 (2014).
58. Nuñez, J. K., Lee, A. S. Y., Engelman, A. & Doudna, J. A. Integrase-mediated spacer acquisition during CRISPR-Cas adaptive immunity. *Nature* **519,** 193–198 (2015).
59. Myler, L. R. *et al.* Single-molecule imaging reveals the mechanism of Exo1 regulation by single-stranded DNA binding proteins. *Proc. Natl. Acad. Sci.* **113,** e1170–e1179 (2016).
60. Finkelstein, I. J., Visnapuu, M.-L. & Greene, E. C. Single-molecule imaging reveals mechanisms of protein disruption by a DNA translocase. *Nature* **468,** 983–987 (2010).
61. Antos, J. M. *et al.* Site-specific N- and C-terminal labeling of a single polypeptide using sortases of different specificity. *J. Am. Chem. Soc.* **131,** 10800–10801 (2009).
62. Guimaraes, C. P. *et al.* Site-specific C-terminal and internal loop labeling of proteins using sortase-mediated reactions. *Nat. Protoc.* **8,** 1787–1799 (2013).
63. Theile, C. S. *et al.* Site-specific N-terminal labeling of proteins using sortase-mediated reactions. *Nat. Protoc.* **8,** 1800–1807 (2013).
64. Malakhov, M. P. *et al.* SUMO fusions and SUMO-specific protease for efficient expression and purification of proteins. *J. Struct. Funct. Genomics* **5,** 75–86 (2004).
65. Kim, Y., de la Torre, A., Leal, A. A. & Finkelstein, I. J. Efficient modification of λ-DNA substrates for single-molecule studies. *Sci. Rep.* **7,** 2071 (2017).
66. Brown, M. W., de la Torre, A. & Finkelstein, I. J. Inserting Extrahelical Structures into Long DNA Substrates for Single-Molecule Studies of DNA Mismatch Repair. *Methods Enzymol.* (2016).

Main Chapters Literature Cited

1. Flajnik MF, Kasahara M. Origin and evolution of the adaptive immune system: genetic events and selective pressures. Nature reviews Genetics. 2010;11(1):47-59. doi: 10.1038/nrg2703. PubMed PMID: PMC3805090.
2. Goren M, Yosef I, Edgar R, Qimron U. The bacterial CRISPR/Cas system as analog of the mammalian adaptive immune system. RNA biology. 2012;9(5):549-54. Epub 2012/05/23. doi: 10.4161/rna.20177. PubMed PMID: 22614830.
3. Karimi Z, Ahmadi A, Najafi A, Ranjbar R. Bacterial CRISPR Regions: General Features and their Potential for Epidemiological Molecular Typing Studies. The open microbiology journal. 2018;12:59-70. Epub 2018/05/15. doi: 10.2174/1874285801812010059. PubMed PMID: 29755603; PMCID: Pmc5925864.
4. Hille F, Richter H, Wong SP, Bratovic M, Ressel S, Charpentier E. The Biology of CRISPR-Cas: Backward and Forward. Cell. 2018;172(6):1239-59. Epub 2018/03/10. doi: 10.1016/j.cell.2017.11.032. PubMed PMID: 29522745.
5. Gleditzsch D, Pausch P, Muller-Esparza H, Ozcan A, Guo X, Bange G, Randau L. PAM identification by CRISPR-Cas effector complexes: diversified mechanisms and structures. RNA biology. 2018. Epub 2018/08/16. doi: 10.1080/15476286.2018.1504546. PubMed PMID: 30109815.
6. Javed MR, Sadaf M, Ahmed T, Jamil A, Nawaz M, Abbas H, Ijaz A. CRISPR-Cas System: History and Prospects as a Genome Editing Tool in Microorganisms. Current microbiology. 2018. Epub 2018/08/06. doi: 10.1007/s00284-018-1547-4. PubMed PMID: 30078067.
7. Marraffini LA, Sontheimer EJ. CRISPR interference: RNA-directed adaptive immunity in bacteria and archaea. Nature reviews Genetics. 2010;11(3):181-90. doi: 10.1038/nrg2749. PubMed PMID: PMC2928866.
8. Bolotin A, Quinquis B, Sorokin A, Ehrlich SD. Clustered regularly interspaced short palindrome repeats (CRISPRs) have spacers of extrachromosomal origin. Microbiology (Reading, England). 2005;151(Pt 8):2551-61. Epub 2005/08/05. doi: 10.1099/mic.0.28048-0. PubMed PMID: 16079334.
9. Horvath P, Romero DA, Coûté-Monvoisin A-C, Richards M, Deveau H, Moineau S, Boyaval P, Fremaux C, Barrangou R. Diversity, Activity, and Evolution of CRISPR Loci in <em>Streptococcus thermophilus</em&gt. Journal of bacteriology. 2008;190(4):1401.
10. Martynov A, Severinov K, Ispolatov I. Optimal number of spacers in CRISPR arrays. PLoS Computational Biology. 2017;13(12):e1005891. doi: 10.1371/journal.pcbi.1005891. PubMed PMID: PMC5749868.
11. Toms A, Barrangou R. On the global CRISPR array behavior in class I systems. Biology Direct. 2017;12(1):20. doi: 10.1186/s13062-017-0193-2.
12. Gasiunas G, Barrangou R, Horvath P, Siksnys V. Cas9–crRNA ribonucleoprotein complex mediates specific DNA cleavage for adaptive immunity in bacteria. Proceedings of the National Academy of Sciences of the United States of America. 2012;109(39):E2579-E86. doi: 10.1073/pnas.1208507109. PubMed PMID: PMC3465414.
13. Li H. Structural principles of CRISPR RNA processing. Structure (London, England : 1993). 2015;23(1):13-20. doi: 10.1016/j.str.2014.10.006. PubMed PMID: PMC4286480.
14. Nishimasu H, Nureki O. Structures and mechanisms of CRISPR RNA-guided effector nucleases. Current Opinion in Structural Biology. 2017;43:68-78. doi: https://doi.org/10.1016/j.sbi.2016.11.013.
15. Karvelis T, Gasiunas G, Miksys A, Barrangou R, Horvath P, Siksnys V. crRNA and tracrRNA guide Cas9-mediated DNA interference in Streptococcus thermophilus. RNA biology. 2013;10(5):841-51. doi: 10.4161/rna.24203. PubMed PMID: PMC3737341.

16. Makarova KS, Koonin EV. Annotation and Classification of CRISPR-Cas Systems. Methods in molecular biology (Clifton, NJ). 2015;1311:47-75. doi: 10.1007/978-1-4939-2687-9_4. PubMed PMID: PMC5901762.
17. Makarova KS, Wolf YI, Alkhnbashi OS, Costa F, Shah SA, Saunders SJ, Barrangou R, Brouns SJJ, Charpentier E, Haft DH, Horvath P, Moineau S, Mojica FJM, Terns RM, Terns MP, White MF, Yakunin AF, Garrett RA, van der Oost J, Backofen R, Koonin EV. An updated evolutionary classification of CRISPR-Cas systems. Nat Rev Micro. 2015;13(11):722-36. doi: 10.1038/nrmicro3569

http://www.nature.com/nrmicro/journal/v13/n11/abs/nrmicro3569.html#supplementary-information.
18. Barrangou R. Diversity of CRISPR-Cas immune systems and molecular machines. Genome Biology. 2015;16:247. doi: 10.1186/s13059-015-0816-9. PubMed PMID: PMC4638107.
19. Shmakov S, Smargon A, Scott D, Cox D, Pyzocha N, Yan W, Abudayyeh OO, Gootenberg JS, Makarova KS, Wolf YI, Severinov K, Zhang F, Koonin EV. Diversity and evolution of class 2 CRISPR–Cas systems. Nature Reviews Microbiology. 2017;15:169. doi: 10.1038/nrmicro.2016.184

https://www.nature.com/articles/nrmicro.2016.184#supplementary-information.
20. Jansen R, Embden JDAv, Gaastra W, Schouls LM. Identification of genes that are associated with DNA repeats in prokaryotes. Molecular Microbiology. 2002;43(6):1565-75. doi: 10.1046/j.1365-2958.2002.02839.x.
21. Nuñez JK, Kranzusch PJ, Noeske J, Wright AV, Davies CW, Doudna JA. Cas1–Cas2 complex formation mediates spacer acquisition during CRISPR–Cas adaptive immunity. Nature structural & molecular biology. 2014;21(6):528-34. doi: 10.1038/nsmb.2820. PubMed PMID: PMC4075942.
22. Nuñez JK, Lee ASY, Engelman A, Doudna JA. Integrase-mediated spacer acquisition during CRISPR–Cas adaptive immunity. Nature. 2015;519(7542):193-8. doi: 10.1038/nature14237. PubMed PMID: PMC4359072.
23. Xiao Y, Ng S, Nam KH, Ke A. How Type II CRISPR-Cas establish immunity through Cas1-Cas2 mediated spacer integration. Nature. 2017;550(7674):137-41. doi: 10.1038/nature24020. PubMed PMID: PMC5832332.
24. Roberts RJ. How restriction enzymes became the workhorses of molecular biology. Proceedings of the National Academy of Sciences. 2005;102(17):5905.
25. Han W, She Q. Chapter One - CRISPR History: Discovery, Characterization, and Prosperity. In: Torres-Ruiz R, Rodriguez-Perales S, editors. Progress in Molecular Biology and Translational Science: Academic Press; 2017. p. 1-21.
26. Ishino Y, Krupovic M, Forterre P. History of CRISPR-Cas from Encounter with a Mysterious Repeated Sequence to Genome Editing Technology. Journal of bacteriology. 2018;200(7). Epub 2018/01/24. doi: 10.1128/jb.00580-17. PubMed PMID: 29358495; PMCID: Pmc5847661.
27. J. D. Genome-editing revolution: my whirlwind year with CRISPR. Nature. 2015(528):469-71. doi: 10.1038/528469a.
28. Lander ES. The Heroes of CRISPR. Cell. 2016;164(1-2):18-28. Epub 2016/01/16. doi: 10.1016/j.cell.2015.12.041. PubMed PMID: 26771483.
29. Wang R, Li H. The mysterious RAMP proteins and their roles in small RNA-based immunity. Protein Science : A Publication of the Protein Society. 2012;21(4):463-70. doi: 10.1002/pro.2044. PubMed PMID: PMC3375746.
30. Makarova KS, Haft DH, Barrangou R, Brouns SJ, Charpentier E, Horvath P, Moineau S, Mojica FJ, Wolf YI, Yakunin AF, van der Oost J, Koonin EV. Evolution and classification of the CRISPR-Cas systems. Nature reviews Microbiology. 2011;9(6):467-77. Epub 2011/05/10. doi: 10.1038/nrmicro2577. PubMed PMID: 21552286; PMCID: PMC3380444.

31. Ishino Y, Shinagawa H, Makino K, Amemura M, Nakata A. Nucleotide sequence of the iap gene, responsible for alkaline phosphatase isozyme conversion in Escherichia coli, and identification of the gene product. Journal of bacteriology. 1987;169(12):5429-33. PubMed PMID: PMC213968.
32. Mojica FJ, Diez-Villasenor C, Garcia-Martinez J, Soria E. Intervening sequences of regularly spaced prokaryotic repeats derive from foreign genetic elements. Journal of molecular evolution. 2005;60(2):174-82. Epub 2005/03/29. doi: 10.1007/s00239-004-0046-3. PubMed PMID: 15791728.
33. Pourcel C, Salvignol G, Vergnaud G. CRISPR elements in Yersinia pestis acquire new repeats by preferential uptake of bacteriophage DNA, and provide additional tools for evolutionary studies. Microbiology (Reading, England). 2005;151(Pt 3):653-63. Epub 2005/03/11. doi: 10.1099/mic.0.27437-0. PubMed PMID: 15758212.
34. Barrangou R, Fremaux C, Deveau H, Richards M, Boyaval P, Moineau S, Romero DA, Horvath P. CRISPR Provides Acquired Resistance Against Viruses in Prokaryotes. Science (New York, NY). 2007;315(5819):1709.
35. Jiang W, Bikard D, Cox D, Zhang F, Marraffini LA. RNA-guided editing of bacterial genomes using CRISPR-Cas systems. Nature Biotechnology. 2013;31:233. doi: 10.1038/nbt.2508

https://www.nature.com/articles/nbt.2508#supplementary-information.
36. Jinek M, East A, Cheng A, Lin S, Ma E, Doudna J. RNA-programmed genome editing in human cells. eLife. 2013;2:e00471. doi: 10.7554/eLife.00471. PubMed PMID: PMC3557905.
37. Sherkow JS. Patent protection for CRISPR: an ELSI review. Journal of Law and the Biosciences. 2017;4(3):565-76. doi: 10.1093/jlb/lsx036. PubMed PMID: PMC5965580.
38. Jiang F, Doudna JA. CRISPR–Cas9 Structures and Mechanisms. Annual Review of Biophysics. 2017;46(1):505-29. doi: 10.1146/annurev-biophys-062215-010822.
39. Pollack A. A Powerful New Way to Edit DNA. New York Times. 2014.
40. Holt N. How DNA Scissors Can Perform Surgery Diresctly On Your Genes. Popular Science. 2014.
41. Molteni M. CRISPR's Epic Patent Fight Changed the Course of Biology. Wired Magazine. 2018.
42. Robinson T. Rampage is laughably dumb, but at least there's plenty of rampaging. The Verge. 2018.
43. Livni E. A new sci-fi thriller about CRISPR makes gene editing terrifyingly easy to understand. Quartz Magazine. 2017.
44. Boddy J. Jennifer Lopez set to produce NBC bio-terror drama C.R.I.S.P.R. Science Magazine. 2016.
45. Mojica FJ, Diez-Villasenor C, Garcia-Martinez J, Almendros C. Short motif sequences determine the targets of the prokaryotic CRISPR defence system. Microbiology (Reading, England). 2009;155(Pt 3):733-40. Epub 2009/02/28. doi: 10.1099/mic.0.023960-0. PubMed PMID: 19246744.
46. Marraffini LA, Sontheimer EJ. Self versus non-self discrimination during CRISPR RNA-directed immunity. Nature. 2010;463(7280):568-71. Epub 2010/01/15. doi: 10.1038/nature08703. PubMed PMID: 20072129; PMCID: Pmc2813891.
47. Lesnik EA, Freier SM. Relative Thermodynamic Stability of DNA, RNA, and DNA:RNA Hybrid Duplexes: Relationship with Base Composition and Structure. Biochemistry. 1995;34(34):10807-15. doi: 10.1021/bi00034a013.
48. Szczelkun MD, Tikhomirova MS, Sinkunas T, Gasiunas G, Karvelis T, Pschera P, Siksnys V, Seidel R. Direct observation of R-loop formation by single RNA-guided Cas9 and Cascade effector complexes. Proceedings of the National Academy of Sciences. 2014;111(27):9798.
49. Nishimasu H, Ran FA, Hsu PD, Konermann S, Shehata SI, Dohmae N, Ishitani R, Zhang F, Nureki O. Crystal structure of Cas9 in complex with guide RNA and target DNA. Cell. 2014;156(5):935-49. Epub 2014/02/18. doi: 10.1016/j.cell.2014.02.001. PubMed PMID: 24529477; PMCID: Pmc4139937.

50. Mulepati S, Héroux A, Bailey S. Crystal structure of a CRISPR RNA–guided surveillance complex bound to a ssDNA target. Science (New York, NY). 2014;345(6203):1479.
51. Jore MM. Structural basis for CRISPR RNA-guided DNA recognition by Cascade. Nat Struct Mol Biol. 2011;18:529-36.
52. Grissa I, Vergnaud G, Pourcel C. CRISPRFinder: a web tool to identify clustered regularly interspaced short palindromic repeats. Nucleic acids research. 2007;35(Web Server issue):W52-W7. doi: 10.1093/nar/gkm360. PubMed PMID: PMC1933234.
53. Nam KH, Haitjema C, Liu X, Ding F, Wang H, DeLisa MP, Ke A. Cas5d protein processes pre-crRNA and assembles into a Cascade-like interference complex in Subtype I-C/Dvulg CRISPR-Cas system. Structure (London, England : 1993). 2012;20(9):1574-84. doi: 10.1016/j.str.2012.06.016. PubMed PMID: PMC3479641.
54. Jung C, Hawkins JA, Jones SK, Jr., Xiao Y, Rybarski JR, Dillard KE, Hussmann J, Saifuddin FA, Savran CA, Ellington AD, Ke A, Press WH, Finkelstein IJ. Massively Parallel Biophysical Analysis of CRISPR-Cas Complexes on Next Generation Sequencing Chips. Cell. 2017;170(1):35-47.e13. doi: 10.1016/j.cell.2017.05.044.
55. Hayes RP, Xiao Y, Ding F, van Erp PB, Rajashankar K, Bailey S, Wiedenheft B, Ke A. Structural basis for promiscuous PAM recognition in type I-E Cascade from E. coli. Nature. 2016;530(7591):499-503. Epub 2016/02/11. doi: 10.1038/nature16995. PubMed PMID: 26863189.
56. Zhao H, Sheng G, Wang J, Wang M, Bunkoczi G, Gong W, Wei Z, Wang Y. Crystal structure of the RNA-guided immune surveillance Cascade complex in Escherichia coli. Nature. 2014;515:147. doi: 10.1038/nature13733.
57. Xiao Y, Luo M, Dolan AE, Liao M, Ke A. Structure basis for RNA-guided DNA degradation by Cascade and Cas3. Science (New York, NY). 2018;361(6397). Epub 2018/06/09. doi: 10.1126/science.aat0839. PubMed PMID: 29880725.
58. Xiao Y, Luo M, Hayes RP, Kim J, Ng S, Ding F, Liao M, Ke A. Structure basis for directional R-loop formation and substrate handover mechanisms in Type I CRISPR-Cas system. Cell. 2017;170(1):48-60.e11. doi: 10.1016/j.cell.2017.06.012. PubMed PMID: PMC5841471.
59. Hochstrasser ML, Taylor DW, Kornfeld JE, Nogales E, Doudna JA. DNA Targeting by a Minimal CRISPR RNA-Guided Cascade. Molecular cell. 2016;63(5):840-51. doi: 10.1016/j.molcel.2016.07.027. PubMed PMID: PMC5111854.
60. Semenova E, Jore MM, Datsenko KA, Semenova A, Westra ER, Wanner B, van der Oost J, Brouns SJJ, Severinov K. Interference by clustered regularly interspaced short palindromic repeat (CRISPR) RNA is governed by a seed sequence. Proceedings of the National Academy of Sciences. 2011;108(25):10098.
61. Wiedenheft B, van Duijn E, Bultema JB, Waghmare SP, Zhou K, Barendregt A, Westphal W, Heck AJR, Boekema EJ, Dickman MJ, Doudna JA. RNA-guided complex from a bacterial immune system enhances target recognition through seed sequence interactions. Proceedings of the National Academy of Sciences of the United States of America. 2011;108(25):10092-7. doi: 10.1073/pnas.1102716108. PubMed PMID: PMC3121849.
62. Zeng Y, Cui Y, Zhang Y, Zhang Y, Liang M, Chen H, Lan J, Song G, Lou J. The initiation, propagation and dynamics of CRISPR-SpyCas9 R-loop complex. Nucleic acids research. 2018;46(1):350-61. doi: 10.1093/nar/gkx1117. PubMed PMID: PMC5758904.
63. Anders C, Niewoehner O, Duerst A, Jinek M. Structural basis of PAM-dependent target DNA recognition by the Cas9 endonuclease. Nature. 2014;513:569. doi: 10.1038/nature13579

https://www.nature.com/articles/nature13579#supplementary-information.
64. Yamano T, Nishimasu H, Zetsche B, Hirano H, Slaymaker IM, Li Y, Fedorova I, Nakane T, Makarova KS, Koonin EV, Ishitani R, Zhang F, Nureki O. Crystal Structure of Cpf1 in Complex with Guide

RNA and Target DNA. Cell. 2016;165(4):949-62. Epub 2016/04/27. doi: 10.1016/j.cell.2016.04.003. PubMed PMID: 27114038; PMCID: Pmc4899970.
65. Mojica FJM, Díez-Villaseñor C, García-Martínez J, Almendros C. Short motif sequences determine the targets of the prokaryotic CRISPR defence system2009;155(3):733-40. doi: doi:10.1099/mic.0.023960-0.
66. Jinek M, Chylinski K, Fonfara I, Hauer M, Doudna JA, Charpentier E. A Programmable Dual-RNA–Guided DNA Endonuclease in Adaptive Bacterial Immunity. Science (New York, NY). 2012.
67. Westra ER, van Erp PBG, Künne T, Wong SP, Staals RHJ, Seegers CLC, Bollen S, Jore MM, Semenova E, Severinov K, de Vos WM, Dame RT, de Vries R, Brouns SJJ, van der Oost J. CRISPR immunity relies on the consecutive binding and degradation of negatively supercoiled invader DNA by Cascade and Cas3. Molecular Cell. 2012;46(5):595-605. doi: 10.1016/j.molcel.2012.03.018. PubMed PMID: PMC3372689.
68. Pattanayak V, Lin S, Guilinger JP, Ma E, Doudna JA, Liu DR. High-throughput profiling of off-target DNA cleavage reveals RNA-programmed Cas9 nuclease specificity. Nature biotechnology. 2013;31(9):839-43. doi: 10.1038/nbt.2673. PubMed PMID: PMC3782611.
69. Huo Y, Nam KH, Ding F, Lee H, Wu L, Xiao Y, Farchione Jr MD, Zhou S, Rajashankar K, Kurinov I, Zhang R, Ke A. Structures of CRISPR Cas3 offer mechanistic insights into Cascade-activated DNA unwinding and degradation. Nat Struct Mol Biol. 2014;21(9):771-7. doi: 10.1038/nsmb.2875

http://www.nature.com/nsmb/journal/v21/n9/abs/nsmb.2875.html#supplementary-information.
70. Anders C, Niewoehner O, Duerst A, Jinek M. Structural basis of PAM-dependent target DNA recognition by the Cas9 endonuclease. Nature. 2014;513(7519):569-73. doi: 10.1038/nature13579

http://www.nature.com/nature/journal/v513/n7519/abs/nature13579.html#supplementary-information.
71. Sternberg SH, Redding S, Jinek M, Greene EC, Doudna JA. DNA interrogation by the CRISPR RNA-guided endonuclease Cas9. Nature. 2014;507:62. doi: 10.1038/nature13011

https://www.nature.com/articles/nature13011#supplementary-information.
72. Westra ER, Semenova E, Datsenko KA, Jackson RN, Wiedenheft B, Severinov K, Brouns SJJ. Type I-E CRISPR-Cas Systems Discriminate Target from Non-Target DNA through Base Pairing-Independent PAM Recognition. PLoS Genetics. 2013;9(9):e1003742. doi: 10.1371/journal.pgen.1003742. PubMed PMID: PMC3764190.
73. Wang R, Preamplume G, Terns MP, Terns RM, Li H. Interaction of the Cas6 riboendonuclease with CRISPR RNAs: recognition and cleavage. Structure (London, England : 1993). 2011;19(2):257-64. doi: 10.1016/j.str.2010.11.014. PubMed PMID: PMC3154685.
74. Osawa T, Inanaga H, Sato C, Numata T. Crystal Structure of the CRISPR-Cas RNA Silencing Cmr Complex Bound to a Target Analog. Molecular Cell. 2015;58(3):418-30. doi: 10.1016/j.molcel.2015.03.018.
75. Majumdar S, Zhao P, Pfister NT, Compton M, Olson S, Glover CVC, Wells L, Graveley BR, Terns RM, Terns MP. Three CRISPR-Cas immune effector complexes coexist in Pyrococcus furiosus. RNA. 2015;21(6):1147-58.
76. Carte J, Wang R, Li H, Terns RM, Terns MP. Cas6 is an endoribonuclease that generates guide RNAs for invader defense in prokaryotes. Genes & Development. 2008;22(24):3489-96. doi: 10.1101/gad.1742908. PubMed PMID: PMC2607076.
77. Sashital DG, Jinek M, Doudna JA. An RNA-induced conformational change required for CRISPR RNA cleavage by the endoribonuclease Cse3. Nat Struct Mol Biol. 2011;18(6):680-7. Epub 2011/05/17. doi: 10.1038/nsmb.2043. PubMed PMID: 21572442.

78. Hochstrasser ML. CasA mediates Cas3-catalyzed target degradation during CRISPR RNA-guided interference. Proc Natl Acad Sci USA. 2014;111:6618-23.
79. Mulepati S, Bailey S. In Vitro Reconstitution of an Escherichia coli RNA-guided Immune System Reveals Unidirectional, ATP-dependent Degradation of DNA Target. The Journal of Biological Chemistry. 2013;288(31):22184-92. doi: 10.1074/jbc.M113.472233. PubMed PMID: PMC3829311.
80. Sinkunas T, Gasiunas G, Fremaux C, Barrangou R, Horvath P, Siksnys V. Cas3 is a single-stranded DNA nuclease and ATP-dependent helicase in the CRISPR/Cas immune system. The EMBO Journal. 2011;30(7):1335-42. doi: 10.1038/emboj.2011.41. PubMed PMID: PMC3094125.
81. Loeff L, Brouns SJJ, Joo C. Repetitive DNA Reeling by the Cascade-Cas3 Complex in Nucleotide Unwinding Steps. Molecular Cell. 2018;70(3):385-94.e3. doi: 10.1016/j.molcel.2018.03.031.
82. Redding S, Sternberg SH, Marshall M, Gibb B, Bhat P, Guegler CK, Wiedenheft B, Doudna JA, Greene EC. Surveillance and processing of foreign DNA by the Escherichia coli CRISPR-Cas system. Cell. 2015;163(4):854-65. doi: 10.1016/j.cell.2015.10.003. PubMed PMID: PMC4636941.
83. Brown MW, Dillard KE, Xiao Y, Dolan AE, Hernandez ET, Dahlhauser S, Kim Y, Myler LR, Anslyn E, Ke A, Finkelstein I. Assembly and translocation of a CRISPR-Cas primed acquisition complex. bioRxiv. 2017.
84. Staals RHJ, Jackson SA, Biswas A, Brouns SJJ, Brown CM, Fineran PC. Interference-driven spacer acquisition is dominant over naive and primed adaptation in a native CRISPR–Cas system. Nature Communications. 2016;7:12853. doi: 10.1038/ncomms12853

https://www.nature.com/articles/ncomms12853#supplementary-information.
85. Künne T, Kieper SN, Bannenberg JW, Vogel AIM, Miellet WR, Klein M, Depken M, Suarez-Diez M, Brouns SJJ. Cas3-Derived Target DNA Degradation Fragments Fuel Primed CRISPR Adaptation. Molecular Cell. 2016;63(5):852-64. doi: https://doi.org/10.1016/j.molcel.2016.07.011.
86. Sternberg SH, Richter H, Charpentier E, Qimron U. Adaptation in CRISPR-Cas Systems. Mol Cell. 2016;61(6):797-808. Epub 2016/03/08. doi: 10.1016/j.molcel.2016.01.030. PubMed PMID: 26949040.
87. Amitai G, Sorek R. CRISPR–Cas adaptation: insights into the mechanism of action. Nature Reviews Microbiology. 2016;14:67. doi: 10.1038/nrmicro.2015.14.
88. Wiedenheft B, Zhou K, Jinek M, Coyle SM, Ma W, Doudna JA. Structural Basis for DNase Activity of a Conserved Protein Implicated in CRISPR-Mediated Genome Defense. Structure. 2009;17(6):904-12. doi: 10.1016/j.str.2009.03.019.
89. Wang J, Li J, Zhao H, Sheng G, Wang M, Yin M, Wang Y. Structural and Mechanistic Basis of PAM-Dependent Spacer Acquisition in CRISPR-Cas Systems. Cell. 2015;163(4):840-53. doi: https://doi.org/10.1016/j.cell.2015.10.008.
90. Wright AV, Liu J-J, Knott GJ, Doxzen KW, Nogales E, Doudna JA. Structures of the CRISPR genome integration complex. Science (New York, NY). 2017.
91. Díez-Villaseñor C, Guzmán NM, Almendros C, García-Martínez J, Mojica FJM. CRISPR-spacer integration reporter plasmids reveal distinct genuine acquisition specificities among CRISPR-Cas I-E variants of Escherichia coli. RNA biology. 2013;10(5):792-802. doi: 10.4161/rna.24023. PubMed PMID: PMC3737337.
92. Nam KH, Kurinov I, Ke A. Crystal Structure of Clustered Regularly Interspaced Short Palindromic Repeats (CRISPR)-associated Csn2 Protein Revealed Ca(2+)-dependent Double-stranded DNA Binding Activity. The Journal of Biological Chemistry. 2011;286(35):30759-68. doi: 10.1074/jbc.M111.256263. PubMed PMID: PMC3162437.
93. Ka D, Lee H, Jung Y-D, Kim K, Seok C, Suh N, Bae E. Crystal Structure of Streptococcus pyogenes Cas1 and Its Interaction with Csn2 in the Type II CRISPR-Cas System. Structure. 2016;24(1):70-9. doi: https://doi.org/10.1016/j.str.2015.10.019.

94. Heler R, Samai P, Modell JW, Weiner C, Goldberg GW, Bikard D, Marraffini LA. Cas9 specifies functional viral targets during CRISPR-Cas adaptation. Nature. 2015;519(7542):199-202. doi: 10.1038/nature14245. PubMed PMID: PMC4385744.
95. Datsenko KA, Pougach K, Tikhonov A, Wanner BL, Severinov K, Semenova E. Molecular memory of prior infections activates the CRISPR/Cas adaptive bacterial immunity system. Nature Communications. 2012;3:945. doi: 10.1038/ncomms1937

https://www.nature.com/articles/ncomms1937#supplementary-information.
96. Radovčić M, Killelea T, Savitskaya E, Wettstein L, Bolt EL, Ivančić-Baće I. CRISPR–Cas adaptation in Escherichia coli requires RecBCD helicase but not nuclease activity, is independent of homologous recombination, and is antagonized by 5′ ssDNA exonucleases. Nucleic acids research. 2018:gky799-gky. doi: 10.1093/nar/gky799.
97. Smith GR. How RecBCD Enzyme and Chi Promote DNA Break Repair and Recombination: a Molecular Biologist's View. Microbiology and Molecular Biology Reviews. 2012;76(2):217.
98. Cong L, Ran FA, Cox D, Lin S, Barretto R, Habib N, Hsu PD, Wu X, Jiang W, Marraffini L, Zhang F. Multiplex Genome Engineering Using CRISPR/Cas Systems. Science (New York, NY). 2013.
99. Wang H, Yang H, Shivalila CS, Dawlaty MM, Cheng AW, Zhang F, Jaenisch R. One-Step Generation of Mice Carrying Mutations in Multiple Genes by CRISPR/Cas-Mediated Genome Engineering. Cell. 2013;153(4):910-8. doi: https://doi.org/10.1016/j.cell.2013.04.025.
100. Chen JS, Dagdas YS, Kleinstiver BP, Welch MM, Sousa AA, Harrington LB, Sternberg SH, Joung JK, Yildiz A, Doudna JA. Enhanced proofreading governs CRISPR-Cas9 targeting accuracy. Nature. 2017;550(7676):407-10. doi: 10.1038/nature24268. PubMed PMID: PMC5918688.
101. Kleinstiver BP, Pattanayak V, Prew MS, Tsai SQ, Nguyen NT, Zheng Z, Joung JK. High-fidelity CRISPR–Cas9 nucleases with no detectable genome-wide off-target effects. Nature. 2016;529:490. doi: 10.1038/nature16526

https://www.nature.com/articles/nature16526#supplementary-information.
102. Wang S, Ren S, Bai R, Xiao P, Zhou Q, Zhou Y, Zhou Z, Niu Yy, Ji W, Chen Y. No off-target mutations in functional genome regions of a CRISPR/Cas9-generated monkey model of muscular dystrophy. Journal of Biological Chemistry. 2018.
103. Zhang X-H, Tee LY, Wang X-G, Huang Q-S, Yang S-H. Off-target Effects in CRISPR/Cas9-mediated Genome Engineering. Molecular Therapy - Nucleic Acids. 2015;4:e264. doi: https://doi.org/10.1038/mtna.2015.37.
104. Team NE. CRISPR off-targets: a reassessment. Nature Methods. 2018;15:229. doi: 10.1038/nmeth.4664.
105. Schaefer KA, Wu W-H, Colgan DF, Tsang SH, Bassuk AG, Mahajan VB. Unexpected mutations after CRISPR–Cas9 editing in vivo. Nature Methods. 2017;14:547. doi: 10.1038/nmeth.4293

https://www.nature.com/articles/nmeth.4293#supplementary-information.
106. Nutter LMJ, Heaney JD, Lloyd KCK, Murray SA, Seavitt JR, Skarnes WC, Teboul L, Brown SDM, Moore M. Response to “Unexpected mutations after CRISPR–Cas9 editing in vivo”. Nature Methods. 2018;15:235. doi: 10.1038/nmeth.4559.
107. Hay EA, Khalaf AR, Marini P, Brown A, Heath K, Sheppard D, MacKenzie A. An analysis of possible off target effects following CAS9/CRISPR targeted deletions of neuropeptide gene enhancers from the mouse genome. Neuropeptides. 2017;64:101-7. doi: 10.1016/j.npep.2016.11.003. PubMed PMID: PMC5529291.
108. Kosicki M, Tomberg K, Bradley A. Repair of double-strand breaks induced by CRISPR–Cas9 leads to large deletions and complex rearrangements. Nature Biotechnology. 2018;36:765. doi: 10.1038/nbt.4192

https://www.nature.com/articles/nbt.4192#supplementary-information.
109. Chen B, Huang B. Imaging genomic elements in living cells using CRISPR/Cas9. Methods in enzymology. 2014;546:337-54. Epub 2014/11/16. doi: 10.1016/b978-0-12-801185-0.00016-7. PubMed PMID: 25398348.
110. Fujita T, Fujii H. Efficient isolation of specific genomic regions and identification of associated proteins by engineered DNA-binding molecule-mediated chromatin immunoprecipitation (enChIP) using CRISPR. Biochemical and biophysical research communications. 2013;439(1):132-6. Epub 2013/08/15. doi: 10.1016/j.bbrc.2013.08.013. PubMed PMID: 23942116.
111. Gilbert LA, Larson MH, Morsut L, Liu Z, Brar GA, Torres SE, Stern-Ginossar N, Brandman O, Whitehead EH, Doudna JA, Lim WA, Weissman JS, Qi LS. CRISPR-mediated modular RNA-guided regulation of transcription in eukaryotes. Cell. 2013;154(2):442-51. Epub 2013/07/16. doi: 10.1016/j.cell.2013.06.044. PubMed PMID: 23849981; PMCID: Pmc3770145.
112. Shen B, Zhang W, Zhang J, Zhou J, Wang J, Chen L, Wang L, Hodgkins A, Iyer V, Huang X, Skarnes WC. Efficient genome modification by CRISPR-Cas9 nickase with minimal off-target effects. Nat Methods. 2014;11(4):399-402. Epub 2014/03/04. doi: 10.1038/nmeth.2857. PubMed PMID: 24584192.
113. Trevino AE, Zhang F. Genome editing using Cas9 nickases. Methods in enzymology. 2014;546:161-74. Epub 2014/11/16. doi: 10.1016/b978-0-12-801185-0.00008-8. PubMed PMID: 25398340.
114. Ran FA, Hsu PD, Lin CY, Gootenberg JS, Konermann S, Trevino AE, Scott DA, Inoue A, Matoba S, Zhang Y, Zhang F. Double nicking by RNA-guided CRISPR Cas9 for enhanced genome editing specificity. Cell. 2013;154(6):1380-9. Epub 2013/09/03. doi: 10.1016/j.cell.2013.08.021. PubMed PMID: 23992846; PMCID: Pmc3856256.
115. Zhang Y, Heidrich N, Ampattu BJ, Gunderson CW, Seifert HS, Schoen C, Vogel J, Sontheimer EJ. Processing-independent CRISPR RNAs limit natural transformation in Neisseria meningitidis. Mol Cell. 2013;50(4):488-503. Epub 2013/05/28. doi: 10.1016/j.molcel.2013.05.001. PubMed PMID: 23706818; PMCID: Pmc3694421.
116. Hou Z, Zhang Y, Propson NE, Howden SE, Chu L-F, Sontheimer EJ, Thomson JA. Efficient genome engineering in human pluripotent stem cells using Cas9 from Neisseria meningitidis. Proceedings of the National Academy of Sciences of the United States of America. 2013;110(39):15644-9. doi: 10.1073/pnas.1313587110. PubMed PMID: PMC3785731.
117. Lee CM, Cradick TJ, Bao G. The Neisseria meningitidis CRISPR-Cas9 System Enables Specific Genome Editing in Mammalian Cells. Molecular Therapy. 2016;24(3):645-54. doi: 10.1038/mt.2016.8. PubMed PMID: PMC4786937.
118. Zetsche B, Gootenberg JS, Abudayyeh OO, Slaymaker IM, Makarova KS, Essletzbichler P, Volz S, Joung J, van der Oost J, Regev A, Koonin EV, Zhang F. Cpf1 is a single RNA-guided endonuclease of a Class 2 CRISPR-Cas system. Cell. 2015;163(3):759-71. doi: 10.1016/j.cell.2015.09.038. PubMed PMID: PMC4638220.
119. Fonfara I, Richter H, Bratovič M, Le Rhun A, Charpentier E. The CRISPR-associated DNA-cleaving enzyme Cpf1 also processes precursor CRISPR RNA. Nature. 2016;532:517. doi: 10.1038/nature17945

https://www.nature.com/articles/nature17945#supplementary-information.
120. Chen JS, Ma E, Harrington LB, Da Costa M, Tian X, Palefsky JM, Doudna JA. CRISPR-Cas12a target binding unleashes indiscriminate single-stranded DNase activity. Science (New York, NY). 2018;360(6387):436.
121. Bloom JD, Arnold FH. In the light of directed evolution: pathways of adaptive protein evolution. Proc Natl Acad Sci U S A. 2009;106 Suppl 1:9995-10000. Epub 2009/06/17. doi: 10.1073/pnas.0901522106. PubMed PMID: 19528653; PMCID: Pmc2702793.

122. Hu JH, Miller SM, Geurts MH, Tang W, Chen L, Sun N, Zeina CM, Gao X, Rees HA, Lin Z, Liu DR. Evolved Cas9 variants with broad PAM compatibility and high DNA specificity. Nature. 2018;556(7699):57-63. Epub 2018/03/08. doi: 10.1038/nature26155. PubMed PMID: 29512652; PMCID: Pmc5951633.
123. Abudayyeh OO, Gootenberg JS, Konermann S, Joung J, Slaymaker IM, Cox DBT, Shmakov S, Makarova KS, Semenova E, Minakhin L, Severinov K, Regev A, Lander ES, Koonin EV, Zhang F. C2c2 is a single-component programmable RNA-guided RNA-targeting CRISPR effector. Science (New York, NY). 2016.
124. Gootenberg JS, Abudayyeh OO, Lee JW, Essletzbichler P, Dy AJ, Joung J, Verdine V, Donghia N, Daringer NM, Freije CA, Myhrvold C, Bhattacharyya RP, Livny J, Regev A, Koonin EV, Hung DT, Sabeti PC, Collins JJ, Zhang F. Nucleic acid detection with CRISPR-Cas13a/C2c2. Science (New York, NY). 2017;356(6336):438-42. Epub 2017/04/15. doi: 10.1126/science.aam9321. PubMed PMID: 28408723; PMCID: Pmc5526198.
125. Gootenberg JS, Abudayyeh OO, Kellner MJ, Joung J, Collins JJ, Zhang F. Multiplexed and portable nucleic acid detection platform with Cas13, Cas12a, and Csm6. Science (New York, NY). 2018;360(6387):439-44. Epub 2018/02/17. doi: 10.1126/science.aaq0179. PubMed PMID: 29449508; PMCID: Pmc5961727.
126. Li L, Li S, Wang J. CRISPR-Cas12b-assisted nucleic acid detection platform. bioRxiv. 2018.
127. Li S-Y, Cheng Q-X, Wang J-M, Li X-Y, Zhang Z-L, Gao S, Cao R-B, Zhao G-P, Wang J. CRISPR-Cas12a-assisted nucleic acid detection. Cell Discovery. 2018;4(1):20. doi: 10.1038/s41421-018-0028-z.
128. La Russa MF, Qi LS. The New State of the Art: Cas9 for Gene Activation and Repression. Molecular and cellular biology. 2015;35(22):3800-9. Epub 2015/09/16. doi: 10.1128/mcb.00512-15. PubMed PMID: 26370509; PMCID: Pmc4609748.
129. Putri RR, Chen L. Spatiotemporal control of zebrafish (Danio rerio) gene expression using a light-activated CRISPR activation system. Gene. 2018;677:273-9. Epub 2018/08/05. doi: 10.1016/j.gene.2018.07.077. PubMed PMID: 30077009.
130. Qi LS, Larson MH, Gilbert LA, Doudna JA, Weissman JS, Arkin AP, Lim WA. Repurposing CRISPR as an RNA-guided platform for sequence-specific control of gene expression. Cell. 2013;152(5):1173-83. Epub 2013/03/05. doi: 10.1016/j.cell.2013.02.022. PubMed PMID: 23452860; PMCID: Pmc3664290.
131. Larson MH, Gilbert LA, Wang X, Lim WA, Weissman JS, Qi LS. CRISPR interference (CRISPRi) for sequence-specific control of gene expression. Nature Protocols. 2013;8:2180. doi: 10.1038/nprot.2013.132.
132. Kuscu C, Adli M. CRISPR-Cas9-AID base editor is a powerful gain-of-function screening tool. Nature Methods. 2016;13:983. doi: 10.1038/nmeth.4076.
133. Kim YB, Komor AC, Levy JM, Packer MS, Zhao KT, Liu DR. Increasing the genome-targeting scope and precision of base editing with engineered Cas9-cytidine deaminase fusions. Nat Biotechnol. 2017;35(4):371-6. Epub 2017/02/14. doi: 10.1038/nbt.3803. PubMed PMID: 28191901; PMCID: Pmc5388574.
134. Ma Y, Zhang J, Yin W, Zhang Z, Song Y, Chang X. Targeted AID-mediated mutagenesis (TAM) enables efficient genomic diversification in mammalian cells. Nat Methods. 2016;13(12):1029-35. Epub 2016/11/01. doi: 10.1038/nmeth.4027. PubMed PMID: 27723754.
135. Nishida K, Arazoe T, Yachie N, Banno S, Kakimoto M, Tabata M, Mochizuki M, Miyabe A, Araki M, Hara KY, Shimatani Z, Kondo A. Targeted nucleotide editing using hybrid prokaryotic and vertebrate adaptive immune systems. Science (New York, NY). 2016;353(6305). Epub 2016/08/06. doi: 10.1126/science.aaf8729. PubMed PMID: 27492474.
136. Krokan HE, Drabløs F, Slupphaug G. Uracil in DNA – occurrence, consequences and repair. Oncogene. 2002;21:8935. doi: 10.1038/sj.onc.1205996.

137. Tantawy AAG. Molecular genetics of hemophilia A: Clinical perspectives. Egyptian Journal of Medical Human Genetics. 2010;11(2):105-14. doi: https://doi.org/10.1016/j.ejmhg.2010.10.005.
138. Williams TN, Thein SL. Sickle Cell Anemia and Its Phenotypes. Annual Review of Genomics and Human Genetics. 2018;19(1):113-47. doi: 10.1146/annurev-genom-083117-021320.
139. Razzouk S. CRISPR-Cas9: A cornerstone for the evolution of precision medicine. Annals of Human Genetics. 2018;82(6):331-57. doi: 10.1111/ahg.12271.
140. Yuan J, Ma Y, Huang T, Chen Y, Peng Y, Li B, Li J, Zhang Y, Song B, Sun X, Ding Q, Song Y, Chang X. Genetic Modulation of RNA Splicing with a CRISPR-Guided Cytidine Deaminase. Mol Cell. 2018;72(2):380-94.e7. Epub 2018/10/09. doi: 10.1016/j.molcel.2018.09.002. PubMed PMID: 30293782.
141. Trounson A, McDonald C. Stem Cell Therapies in Clinical Trials: Progress and Challenges. Cell Stem Cell. 2015;17(1):11-22. Epub 2015/07/04. doi: 10.1016/j.stem.2015.06.007. PubMed PMID: 26140604.
142. Yu VWC, Liu Y, Curran M, Zhang P, Snead J, Schmedt C, Yang Y, Lin VG, Tschantz WR, Quinn L, Russ C, Clarkson S, Janiak A, Stewart M, Mulumba Y, Lescarbeau R, Murray B, Seitzer J, Strapps W, Huang H-R, Sloan K, Mickanin CS, Klickstein L, Stevenson S. CRISPR/Cas9 Gene-Edited Hematopoietic Stem Cell Therapy for Sickle Cell Disease. Blood. 2017;130(Suppl 1):535.
143. Dever DP, Bak RO, Reinisch A, Camarena J, Washington G, Nicolas CE, Pavel-Dinu M, Saxena N, Wilkens AB, Mantri S, Uchida N, Hendel A, Narla A, Majeti R, Weinberg KI, Porteus MH. CRISPR/Cas9 Beta-globin Gene Targeting in Human Hematopoietic Stem Cells. Nature. 2016;539(7629):384-9. doi: 10.1038/nature20134. PubMed PMID: PMC5898607.
144. Hou P, Chen S, Wang S, Yu X, Chen Y, Jiang M, Zhuang K, Ho W, Hou W, Huang J, Guo D. Genome editing of CXCR4 by CRISPR/cas9 confers cells resistant to HIV-1 infection. Scientific Reports. 2015;5:15577. doi: 10.1038/srep15577. PubMed PMID: PMC4612538.
145. Park C-Y, Kim Duk H, Son Jeong S, Sung Jin J, Lee J, Bae S, Kim J-H, Kim D-W, Kim J-S. Functional Correction of Large Factor VIII Gene Chromosomal Inversions in Hemophilia A Patient-Derived iPSCs Using CRISPR-Cas9. Cell Stem Cell. 2015;17(2):213-20. doi: 10.1016/j.stem.2015.07.001.
146. Xu L, Yang H, Gao Y, Chen Z, Xie L, Liu Y, Liu Y, Wang X, Li H, Lai W, He Y, Yao A, Ma L, Shao Y, Zhang B, Wang C, Chen H, Deng H. CRISPR/Cas9-Mediated CCR5 Ablation in Human Hematopoietic Stem/Progenitor Cells Confers HIV-1 Resistance In Vivo. Molecular Therapy. 2017;25(8):1782-9. doi: 10.1016/j.ymthe.2017.04.027. PubMed PMID: PMC5542791.
147. Xu X, Tay Y, Sim B, Yoon S-I, Huang Y, Ooi J, Utami KH, Ziaei A, Ng B, Radulescu C, Low D, Ng AYJ, Loh M, Venkatesh B, Ginhoux F, Augustine GJ, Pouladi MA. Reversal of Phenotypic Abnormalities by CRISPR/Cas9-Mediated Gene Correction in Huntington Disease Patient-Derived Induced Pluripotent Stem Cells. Stem Cell Reports. 2017;8(3):619-33. doi: 10.1016/j.stemcr.2017.01.022.
148. Wiles MV, Qin W, Cheng AW, Wang H. CRISPR-Cas9-mediated genome editing and guide RNA design. Mammalian genome : official journal of the International Mammalian Genome Society. 2015;26(9-10):501-10. Epub 05/20. doi: 10.1007/s00335-015-9565-z. PubMed PMID: 25991564.
149. Zhang L, Jia R, Palange NJ, Satheka AC, Togo J, An Y, Humphrey M, Ban L, Ji Y, Jin H, Feng X, Zheng Y. Large genomic fragment deletions and insertions in mouse using CRISPR/Cas9. PloS one. 2015;10(3):e0120396-e. doi: 10.1371/journal.pone.0120396. PubMed PMID: 25803037.
150. Song Y, Lai L, Li Z. Large-scale genomic deletions mediated by CRISPR/Cas9 system. Oncotarget. 2017;8(4):5647-. doi: 10.18632/oncotarget.14543. PubMed PMID: 28077794.
151. Song Y, Yuan L, Wang Y, Chen M, Deng J, Lv Q, Sui T, Li Z, Lai L. Efficient dual sgRNA-directed large gene deletion in rabbit with CRISPR/Cas9 system. Cellular and Molecular Life Sciences. 2016;73(15):2959-68. doi: 10.1007/s00018-016-2143-z.
152. Cox C, Bignell G, Greenman C, Stabenau A, Warren W, Stephens P, Davies H, Watt S, Teague J, Edkins S, Birney E, Easton DF, Wooster R, Futreal PA, Stratton MR. A survey of homozygous deletions in

human cancer genomes. Proceedings of the National Academy of Sciences of the United States of America. 2005;102(12):4542.
153. Pylkäs K, Erkko H, Nikkilä J, Sólyom S, Winqvist R. Analysis of large deletions in BRCA1, BRCA2 and PALB2 genes in Finnish breast and ovarian cancer families. BMC cancer. 2008;8:146-. doi: 10.1186/1471-2407-8-146. PubMed PMID: 18501021.
154. Bignell GR, Greenman CD, Davies H, Butler AP, Edkins S, Andrews JM, Buck G, Chen L, Beare D, Latimer C, Widaa S, Hinton J, Fahey C, Fu B, Swamy S, Dalgliesh GL, Teh BT, Deloukas P, Yang F, Campbell PJ, Futreal PA, Stratton MR. Signatures of mutation and selection in the cancer genome. Nature. 2010;463(7283):893-8. doi: 10.1038/nature08768. PubMed PMID: 20164919.
155. Rath D, Amlinger L, Hoekzema M, Devulapally PR, Lundgren M. Efficient programmable gene silencing by Cascade. Nucleic acids research. 2015;43(1):237-46. doi: 10.1093/nar/gku1257.
156. Luo ML, Mullis AS, Leenay RT, Beisel CL. Repurposing endogenous type I CRISPR-Cas systems for programmable gene repression. Nucleic acids research. 2015;43(1):674-81. doi: 10.1093/nar/gku971. PubMed PMID: PMC4288209.
157. Li Y, Pan S, Zhang Y, Ren M, Feng M, Peng N, Chen L, Liang YX, She Q. Harnessing Type I and Type III CRISPR-Cas systems for genome editing. Nucleic acids research. 2016;44(4):e34-e. doi: 10.1093/nar/gkv1044. PubMed PMID: PMC4770200.
158. Hayes RP, Xiao Y, Ding F, van Erp PBG, Rajashankar K, Bailey S, Wiedenheft B, Ke A. Structural basis for promiscuous PAM recognition in Type I-E Cascade from E. coli. Nature. 2016;530(7591):499-503. doi: 10.1038/nature16995. PubMed PMID: PMC5134256.
159. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W, Funke R, Gage D, Harris K, Heaford A, Howland J, Kann L, Lehoczky J, LeVine R, McEwan P, McKernan K, Meldrim J, Mesirov JP, Miranda C, Morris W, Naylor J, Raymond C, Rosetti M, Santos R, Sheridan A, Sougnez C, Stange-Thomann Y, Stojanovic N, Subramanian A, Wyman D, Rogers J, Sulston J, Ainscough R, Beck S, Bentley D, Burton J, Clee C, Carter N, Coulson A, Deadman R, Deloukas P, Dunham A, Dunham I, Durbin R, French L, Grafham D, Gregory S, Hubbard T, Humphray S, Hunt A, Jones M, Lloyd C, McMurray A, Matthews L, Mercer S, Milne S, Mullikin JC, Mungall A, Plumb R, Ross M, Shownkeen R, Sims S, Waterston RH, Wilson RK, Hillier LW, McPherson JD, Marra MA, Mardis ER, Fulton LA, Chinwalla AT, Pepin KH, Gish WR, Chissoe SL, Wendl MC, Delehaunty KD, Miner TL, Delehaunty A, Kramer JB, Cook LL, Fulton RS, Johnson DL, Minx PJ, Clifton SW, Hawkins T, Branscomb E, Predki P, Richardson P, Wenning S, Slezak T, Doggett N, Cheng JF, Olsen A, Lucas S, Elkin C, Uberbacher E, Frazier M, Gibbs RA, Muzny DM, Scherer SE, Bouck JB, Sodergren EJ, Worley KC, Rives CM, Gorrell JH, Metzker ML, Naylor SL, Kucherlapati RS, Nelson DL, Weinstock GM, Sakaki Y, Fujiyama A, Hattori M, Yada T, Toyoda A, Itoh T, Kawagoe C, Watanabe H, Totoki Y, Taylor T, Weissenbach J, Heilig R, Saurin W, Artiguenave F, Brottier P, Bruls T, Pelletier E, Robert C, Wincker P, Smith DR, Doucette-Stamm L, Rubenfield M, Weinstock K, Lee HM, Dubois J, Rosenthal A, Platzer M, Nyakatura G, Taudien S, Rump A, Yang H, Yu J, Wang J, Huang G, Gu J, Hood L, Rowen L, Madan A, Qin S, Davis RW, Federspiel NA, Abola AP, Proctor MJ, Myers RM, Schmutz J, Dickson M, Grimwood J, Cox DR, Olson MV, Kaul R, Raymond C, Shimizu N, Kawasaki K, Minoshima S, Evans GA, Athanasiou M, Schultz R, Roe BA, Chen F, Pan H, Ramser J, Lehrach H, Reinhardt R, McCombie WR, de la Bastide M, Dedhia N, Blocker H, Hornischer K, Nordsiek G, Agarwala R, Aravind L, Bailey JA, Bateman A, Batzoglou S, Birney E, Bork P, Brown DG, Burge CB, Cerutti L, Chen HC, Church D, Clamp M, Copley RR, Doerks T, Eddy SR, Eichler EE, Furey TS, Galagan J, Gilbert JG, Harmon C, Hayashizaki Y, Haussler D, Hermjakob H, Hokamp K, Jang W, Johnson LS, Jones TA, Kasif S, Kaspryzk A, Kennedy S, Kent WJ, Kitts P, Koonin EV, Korf I, Kulp D, Lancet D, Lowe TM, McLysaght A, Mikkelsen T, Moran JV, Mulder N, Pollara VJ, Ponting CP, Schuler G, Schultz J, Slater G, Smit AF, Stupka E, Szustakowki J, Thierry-Mieg D, Thierry-Mieg J, Wagner L, Wallis J, Wheeler R, Williams A, Wolf YI, Wolfe KH, Yang SP, Yeh RF, Collins F, Guyer MS, Peterson J, Felsenfeld A, Wetterstrand KA, Patrinos A, Morgan MJ, de Jong P, Catanese JJ, Osoegawa K, Shizuya H, Choi S, Chen YJ, Szustakowki J. Initial sequencing and analysis of

the human genome. Nature. 2001;409(6822):860-921. Epub 2001/03/10. doi: 10.1038/35057062. PubMed PMID: 11237011.
160. Kleinstiver BP, Prew MS, Tsai SQ, Topkar VV, Nguyen NT, Zheng Z, Gonzales APW, Li Z, Peterson RT, Yeh J-RJ, Aryee MJ, Joung JK. Engineered CRISPR-Cas9 nucleases with altered PAM specificities. Nature. 2015;523:481. doi: 10.1038/nature14592

https://www.nature.com/articles/nature14592#supplementary-information.
161. Terns MP, Terns RM. CRISPR-based adaptive immune systems. Curr Opin Microbiol. 2011;14:321-7.
162. Jore MM, Brouns SJ, van der Oost J. RNA in defense: CRISPRs protect prokaryotes against mobile genetic elements. Cold Spring Harb Perspect Biol. 2012;4:a003657.
163. Wiedenheft B, Sternberg SH, Doudna JA. RNA-guided genetic silencing systems in bacteria and archaea. Nature. 2012;482:331-8.
164. Marraffini LA, Sontheimer EJ. CRISPR interference limits horizontal gene transfer in staphylococci by targeting DNA. Science (New York, NY). 2008;322:1843-5.
165. Makarova KS. Evolution and classification of the CRISPR-Cas systems. Nat Rev Microbiol. 2011;9:467-77.
166. Wiedenheft B. Structures of the RNA-guided surveillance complex from a bacterial immune system. Nature. 2011;477:486-9.
167. Westra ER. CRISPR immunity relies on the consecutive binding and degradation of negatively supercoiled invader DNA by Cascade and Cas3. Mol Cell. 2012;46:595-605.
168. Brouns SJ. Small CRISPR RNAs guide antiviral defense in prokaryotes. Science (New York, NY). 2008;321:960-4.
169. Jackson RN, Lavin M, Carter J, Wiedenheft B. Fitting CRISPR-associated Cas3 into the helicase family tree. Curr Opin Struct Biol. 2014;24:106-14.
170. Mulepati S, Bailey S. In vitro reconstitution of an Escherichia coli RNA-guided immune system reveals unidirectional, ATP-dependent degradation of DNA target. J Biol Chem. 2013;288:22184-92.
171. Huo Y, Nam KH, Ding F, Lee H, Wu L, Xiao Y, Farchione FD, Zhou S, Rajashankar R, Kurinov I, Zhang R, Ke A. Structures of CRISPR Cas3 offer mechanistic insights into Cascade-activated DNA unwinding and degradation. Nature structural & molecular biology. 2014;21(9):771-7. doi: 10.1038/nsmb.2875. PubMed PMID: PMC4156918.
172. Tay M, Liu S, Yuan YA. Crystal structure of Thermobifida fusca Cse1 reveals target DNA binding site. Protein science : a publication of the Protein Society. 2015;24(2):236-45. Epub 2014/11/26. doi: 10.1002/pro.2609. PubMed PMID: 25420472; PMCID: Pmc4315661.
173. Barrangou R, Fremaux C, Deveau H, Richards M, Boyaval P, Moineau S, Romero DA, Horvath P. CRISPR provides acquired resistance against viruses in prokaryotes. Science. 2007;315(5819):1709-12.
174. Bolotin A, Quinquis B, Sorokin A, Ehrlich SD. Clustered regularly interspaced short palindrome repeats (CRISPRs) have spacers of extrachromosomal origin. Microbiology. 2005;151(8):2551-61.
175. Mojica FJ, García-Martínez J, Soria E. Intervening sequences of regularly spaced prokaryotic repeats derive from foreign genetic elements. Journal of molecular evolution. 2005;60(2):174-82.
176. Marraffini LA, Sontheimer EJ. CRISPR interference limits horizontal gene transfer in staphylococci by targeting DNA. Science. 2008;322(5909):1843-5. doi: 10.1126/science.1165771. PubMed PMID: 19095942; PMCID: PMC2695655.
177. Makarova KS, Grishin NV, Shabalina SA, Wolf YI, Koonin EV. A putative RNA-interference-based immune system in prokaryotes: computational analysis of the predicted enzymatic machinery, functional analogies with eukaryotic RNAi, and hypothetical mechanisms of action. Biol Direct. 2006;1:7. doi: 10.1186/1745-6150-1-7. PubMed PMID: 16545108; PMCID: PMC1462988.

178. Makarova KS, Koonin EV. Annotation and Classification of CRISPR-Cas Systems. Methods Mol Biol. 2015;1311:47-75. doi: 10.1007/978-1-4939-2687-9_4. PubMed PMID: 25981466.
179. Shmakov S, Abudayyeh OO, Makarova KS, Wolf YI, Gootenberg JS, Semenova E, Minakhin L, Joung J, Konermann S, Severinov K, Zhang F, Koonin EV. Discovery and Functional Characterization of Diverse Class 2 CRISPR-Cas Systems. Mol Cell. 2015;60(3):385-97. doi: 10.1016/j.molcel.2015.10.008. PubMed PMID: 26593719; PMCID: PMC4660269.
180. Jinek M, Chylinski K, Fonfara I, Hauer M, Doudna JA, Charpentier E. A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. Science. 2012;337(6096):816-21. Epub 2012/06/30. doi: 10.1126/science.1225829. PubMed PMID: 22745249.
181. Cong L, Ran FA, Cox D, Lin S, Barretto R, Habib N, Hsu PD, Wu X, Jiang W, Marraffini LA, Zhang F. Multiplex genome engineering using CRISPR/Cas systems. Science. 2013;339(6121):819-23. Epub 2013/01/05. doi: 10.1126/science.1231143. PubMed PMID: 23287718; PMCID: PMC3795411.
182. Mali P, Yang L, Esvelt KM, Aach J, Guell M, DiCarlo JE, Norville JE, Church GM. RNA-guided human genome engineering via Cas9. Science. 2013;339(6121):823-6. Epub 2013/01/05. doi: 10.1126/science.1232033. PubMed PMID: 23287722; PMCID: PMC3712628.
183. Komor AC, Badran AH, Liu DR. CRISPR-Based Technologies for the Manipulation of Eukaryotic Genomes. Cell. 2017;168(1-2):20-36. Epub 2016/11/22. doi: 10.1016/j.cell.2016.10.044. PubMed PMID: 27866654; PMCID: PMC5235943.
184. Brouns SJ, Jore MM, Lundgren M, Westra ER, Slijkhuis RJ, Snijders AP, Dickman MJ, Makarova KS, Koonin EV, van der Oost J. Small CRISPR RNAs guide antiviral defense in prokaryotes. Science. 2008;321(5891):960-4. Epub 2008/08/16. doi: 10.1126/science.1159689. PubMed PMID: 18703739.
185. Wiedenheft B, Lander GC, Zhou K, Jore MM, Brouns SJ, van der Oost J, Doudna JA, Nogales E. Structures of the RNA-guided surveillance complex from a bacterial immune system. Nature. 2011;477(7365):486-9. Epub 2011/09/23. doi: 10.1038/nature10402. PubMed PMID: 21938068.
186. Westra ER, van Erp PB, Kunne T, Wong SP, Staals RH, Seegers CL, Bollen S, Jore MM, Semenova E, Severinov K, de Vos WM, Dame RT, de Vries R, Brouns SJ, van der Oost J. CRISPR immunity relies on the consecutive binding and degradation of negatively supercoiled invader DNA by Cascade and Cas3. Mol Cell. 2012;46(5):595-605. Epub 2012/04/24. doi: 10.1016/j.molcel.2012.03.018. PubMed PMID: 22521689; PMCID: 3372689.
187. Sinkunas T, Gasiunas G, Waghmare SP, Dickman MJ, Barrangou R, Horvath P, Siksnys V. In vitro reconstitution of Cascade-mediated CRISPR immunity in Streptococcus thermophilus. The EMBO journal. 2013;32(3):385-94. Epub 2013/01/22. doi: 10.1038/emboj.2012.352. PubMed PMID: 23334296; PMCID: 3567492.
188. Mulepati S, Bailey S. In vitro reconstitution of an Escherichia coli RNA-guided immune system reveals unidirectional, ATP-dependent degradation of DNA target. The Journal of biological chemistry. 2013;288(31):22184-92. Epub 2013/06/14. doi: 10.1074/jbc.M113.472233. PubMed PMID: 23760266; PMCID: 3829311.
189. Hochstrasser ML, Taylor DW, Bhat P, Guegler CK, Sternberg SH, Nogales E, Doudna JA. CasA mediates Cas3-catalyzed target degradation during CRISPR RNA-guided interference. Proceedings of the National Academy of Sciences of the United States of America. 2014;111(18):6618-23. Epub 2014/04/22. doi: 10.1073/pnas.1405079111. PubMed PMID: 24748111; PMCID: 4020112.
190. Jackson RN, Golden SM, van Erp PB, Carter J, Westra ER, Brouns SJ, van der Oost J, Terwilliger TC, Read RJ, Wiedenheft B. Crystal structure of the CRISPR RNA-guided surveillance complex from Escherichia coli. Science. 2014;345(6203):1473-9. doi: 10.1126/science.1256328. PubMed PMID: 25103409; PMCID: 4188430.
191. Zhao H, Sheng G, Wang J, Wang M, Bunkoczi G, Gong W, Wei Z, Wang Y. Crystal structure of the RNA-guided immune surveillance Cascade complex in Escherichia coli. Nature. 2014;515(7525):147-50. doi: 10.1038/nature13733. PubMed PMID: 25118175.

192. Mulepati S, Heroux A, Bailey S. Crystal structure of a CRISPR RNA-guided surveillance complex bound to a ssDNA target. Science. 2014;345(6203):1479-84. doi: 10.1126/science.1256996. PubMed PMID: 25123481.
193. Rutkauskas M, Sinkunas T, Songailiene I, Tikhomirova MS, Siksnys V, Seidel R. Directional R-Loop Formation by the CRISPR-Cas Surveillance Complex Cascade Provides Efficient Off-Target Site Rejection. Cell reports. 2015. doi: 10.1016/j.celrep.2015.01.067. PubMed PMID: 25753419.
194. Huo Y, Nam KH, Ding F, Lee H, Wu L, Xiao Y, Farchione MD, Jr., Zhou S, Rajashankar K, Kurinov I, Zhang R, Ke A. Structures of CRISPR Cas3 offer mechanistic insights into Cascade-activated DNA unwinding and degradation. Nature structural & molecular biology. 2014;21(9):771-7. doi: 10.1038/nsmb.2875. PubMed PMID: 25132177; PMCID: 4156918.
195. Xiao Y, Luo M, Hayes RP, Kim J, Ng S, Ding F, Liao M, Ke A. Structure Basis for Directional R-loop Formation and Substrate Handover Mechanisms in Type I CRISPR-Cas System. Cell. 2017;170(1):48-60 e11. doi: 10.1016/j.cell.2017.06.012. PubMed PMID: 28666122.
196. Mulepati S, Bailey S. Structural and Biochemical Analysis of Nuclease Domain of Clustered Regularly Interspaced Short Palindromic Repeat (CRISPR)-associated Protein 3 (Cas3). The Journal of Biological Chemistry. 2011;286(36):31896-903. doi: 10.1074/jbc.M111.270017. PubMed PMID: PMC3173111.
197. Redding S, Sternberg SH, Marshall M, Gibb B, Bhat P, Guegler CK, Wiedenheft B, Doudna JA, Greene EC. Surveillance and Processing of Foreign DNA by the Escherichia coli CRISPR-Cas System. Cell. 2015;163(4):854-65. doi: 10.1016/j.cell.2015.10.003. PubMed PMID: 26522594; PMCID: PMC4636941.
198. Diao Y, Fang R, Li B, Meng Z, Yu J, Qiu Y, Lin KC, Huang H, Liu T, Marina RJ, Jung I, Shen Y, Guan KL, Ren B. A tiling-deletion-based genetic screen for cis-regulatory element identification in mammalian cells. Nat Methods. 2017;14(6):629-35. Epub 2017/04/19. doi: 10.1038/nmeth.4264. PubMed PMID: 28417999; PMCID: PMC5490986.
199. Shalem O, Sanjana NE, Hartenian E, Shi X, Scott DA, Mikkelson T, Heckl D, Ebert BL, Root DE, Doench JG, Zhang F. Genome-scale CRISPR-Cas9 knockout screening in human cells. Science. 2014;343(6166):84-7. Epub 2013/12/18. doi: 10.1126/science.1247005. PubMed PMID: 24336571; PMCID: PMC4089965.
200. Fulco CP, Munschauer M, Anyoha R, Munson G, Grossman SR, Perez EM, Kane M, Cleary B, Lander ES, Engreitz JM. Systematic mapping of functional enhancer-promoter connections with CRISPR interference. Science. 2016;354(6313):769-73. Epub 2016/10/07. doi: 10.1126/science.aag2445. PubMed PMID: 27708057; PMCID: PMC5438575.
201. Sanjana NE, Wright J, Zheng K, Shalem O, Fontanillas P, Joung J, Cheng C, Regev A, Zhang F. High-resolution interrogation of functional elements in the noncoding genome. Science. 2016;353(6307):1545-9. Epub 2016/10/07. doi: 10.1126/science.aaf7613. PubMed PMID: 27708104; PMCID: PMC5144102.
202. Picelli S, Bjorklund AK, Reinius B, Sagasser S, Winberg G, Sandberg R. Tn5 transposase and tagmentation procedures for massively scaled sequencing projects. Genome Res. 2014;24(12):2033-40. Epub 2014/08/01. doi: 10.1101/gr.177881.114. PubMed PMID: 25079858; PMCID: PMC4248319.
203. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics. 2014;30(15):2114-20. Epub 2014/04/04. doi: 10.1093/bioinformatics/btu170. PubMed PMID: 24695404; PMCID: PMC4103590.
204. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Methods. 2012;9(4):357-9. Epub 2012/03/06. doi: 10.1038/nmeth.1923. PubMed PMID: 22388286; PMCID: PMC3322381.
205. Nuñez JK, Harrington LB, Kranzusch PJ, Engelman AN, Doudna JA. Foreign DNA capture during CRISPR-Cas adaptive immunity. Nature. 2015;527(7579):535-8. Epub 10/21. doi: 10.1038/nature15760. PubMed PMID: 26503043.

206. Globyte V, Lee SH, Bae T, Kim J-S, Joo C. CRISPR Cas9 searches for a protospacer adjacent motif by one-dimensional diffusion. bioRxiv. 2018.
207. Xue C, Zhu Y, Zhang X, Shin Y-K, Sashital DG. Real-Time Observation of Target Search by the CRISPR Surveillance Complex Cascade. Cell Reports. 2017;21(13):3717-27. doi: https://doi.org/10.1016/j.celrep.2017.11.110.
208. Xue C, Whitis NR, Sashital DG. Conformational Control of Cascade Interference and Priming Activities in CRISPR Immunity. Molecular Cell. 2016;64(4):826-34. doi: https://doi.org/10.1016/j.molcel.2016.09.033.
209. Krivoy A, Rutkauskas M, Kuznedelov K, Musharova O, Rouillon C, Severinov K, Seidel R. Primed CRISPR adaptation in Escherichia coli cells does not depend on conformational changes in the Cascade effector complex detected in Vitro. Nucleic acids research. 2018;46(8):4087-98. Epub 2018/03/30. doi: 10.1093/nar/gky219. PubMed PMID: 29596641; PMCID: Pmc5934681.